多模态融合的情感识别研究

Research on Multi-modal Fusion Emotion Recognition

学科专业：计算机科学与技术
研 究 生：曹田熠
指导教师：王建荣 副教授

天津大学计算机科学与技术学院
二零一二年十二月

独创性声明

本人声明所呈交的学位论文是本人在导师指导下进行的研究工作和取得的研究成果，除了文中特别加以标注和致谢之处外，论文中不包含其他人已经发表或撰写过的研究成果，也不包含为获得 **天津大学** 或其他教育机构的学位或证书而使用过的材料。与我一同工作的同志对本研究所做的任何贡献均已在论文中作了明确的说明并表示了谢意。

学位论文作者签名：　　签字日期：　　年　　月　　日

学位论文版权使用授权书

本学位论文作者完全了解 **天津大学** 有关保留、使用学位论文的规定。特授权 **天津大学** 可以将学位论文的全部或部分内容编入有关数据库进行检索，并采用影印、缩印或扫描等复制手段保存、汇编以供查阅和借阅。同意学校向国家有关部门或机构送交论文的复印件和磁盘。

（保密的学位论文在解密后适用本授权说明）

学位论文作者签名：　　导师签名：

签字日期：　　年　　月　　日　　签字日期：　　年　　月　　日

摘 要

情感是人们在沟通交流的过程中传递的重要信息，情感状态的变化影响着人们的感知和决策。情感识别是模式识别的重要研究领域，它将情感维度引入人机交互。情感表达的模态包括面部表情、语音、姿势、生理信号、文字等，情感识别本质上是一个多模态融合的问题。

提出一种多模态融合的情感识别算法，从面部图像序列和语音信号中提取表情和语音特征，基于隐马尔可夫模型和多层感知器设计融合表情和语音模态的情感分类器。建立面部表情图像的主动外观模型，实现面部特征点的定位和跟踪；根据面部特征点的位移，计算面部动画参数作为表情特征。对语音信号作时域、和频域分析，提取各帧的短时平均能量、基音频率和共振峰作为语音特征。利用提取的表情和语音特征，采用 Viterbi 算法训练各种表情和语音情感的隐马尔可夫模型；利用特征向量关于各隐马尔可夫模型的条件概率，采用反向传播学习算法训练多层感知器。实验结果表明，融合表情和语音的情感识别算法在识别样本中的高兴、悲伤、愤怒、厌恶等情感状态时具有较高的准确率。

提出的多模态识别算法较好地利用了视频和音频中的情感信息，相比于仅利用语音模态的识别结果有较大的提升，相比于表情模态的识别结果也有一定改进，是一种可以采用的情感识别算法。

关键词： 情感识别 多模态融合 隐马尔可夫模型 人工神经网络

ABSTRACT

Emotion plays an important role in human communications. Changes of affective states have an impact on people's perception and decision. Emotion recognition is a significant research field of pattern recognition. It introduces emotion into human-computer interaction. People express their emotions through facial expressions, speech, postures, physiological signals, characters, and so on. As a result, emotion recognition is inherently a matter of multi-modal fusion.

In this paper, we construct a framework of multi-modal fusion emotion recognition. Facial expression features and speech features are respectively extracted from image sequences and speech signals. An emotion classifier is designed to fuse facial expression and speech modalities based on Hidden Markov Models and Multi-layer Perceptron. In order to locate and track facial feature points, we construct an Active Appearance Model for facial images with all kinds of expressions. Facial Animation Parameters are calculated from motions of facial feature points as expression features. We extract short-term mean energy, fundamental frequency and formant frequencies from each speech frame as speech features. We use expression features and speech features to train Hidden Markov Models based on Viterbi Algorithm. Multi-layer Perceptron which fuses expression and speech modalities is trained based Back-propagation Algorithm. Experiments indicate that multi-modal fusion emotion recognition algorithm which is presented in this paper has relatively high recognition accuracies.

The proposed algorithm makes use of affective information from video and audio. The approach has better performance and robustness than methods using only video or audio separately.

KEY WORDS: Emotion recognition, Multi-modal Fusion, HMM, ANN

目　　录

第一章 绪论

1.1 课题背景

情感是人对客观事物是否满足自己的需要而产生的体验态度，是人们在沟通交流的过程中传递的重要信息。情感状态影响着信息表达的方式和信息传递的效果。情感通常以言语、文字、面部表情、肢体动作等为载体。

情感识别是模式识别的重要研究领域，它将情感维度引入人机交互。情感识别的研究涵盖诸多学科，其相关研究方向包括计算机视觉、语音信号处理、人工智能、心理学、社会学等。情感识别技术在智能人机交互、医疗护理、残疾人辅助、汽车和飞机驾驶、安全、通信等领域有广泛的应用。

情感表达的模态包括面部表情、语音、姿势、生理信号、文字等。其中，面部表情通过采集人脸图像获得，语音情感由带情感的语音信号中提取，肢体动作较直接地表达了人们的情感，利用血压、脉搏、皮肤电、脑电等生理信号可以发现人们刻意掩藏的情感，运用自然语言处理技术则可从字里行间读出作者的情感。在人们对话交流的过程中，面部表情作用最大，有研究认为其对情感表达的贡献超过五成；说话方式即语音情感次之，贡献不到四成；说话内容的作用仅占一成。

心理学家 Ekman 认为，人们比较容易识别的面部表情有六种，分别是高兴、愤怒、悲伤、害怕、惊讶和厌恶。面部表情的产生源于人的心理反应，大脑皮层发送信号，经过神经传至面部肌肉组织，控制肌肉运动，使平静的人脸变形为带有表情的人脸。由于种族文化和地域文化的差异，不同民族、不同地区的人们通过面部表情表达情感的方式有较大的差别。同一种情感状态，对于不同文化中的人们，其表情产生的部位、面部器官形变的轨迹和强度、表情的反应速度和持续时间都会有所差别。即便同一文化中的人们，由于所处具体场景的差异，其情感表达也会有所不同。

表情识别技术应用领域广泛。在人机交互中，人的表情可以作为命令控制计算机执行操作。与传统的利用鼠标、键盘、触摸屏等发送控制命令的方式相比，这种方式更加简便且更符合人的习惯。比如，人们做出微笑的表情，表示确认当前的任务，或做出不满的表情，表示取消当前的任务。

表情识别技术可应用于辅助医疗护理。在病房或老年人居住的房间中安装摄像头，实时采集人脸图像，跟踪看护对象的面部表情。当受看护的病人或老人突然做出剧烈的表情时，设备立即报警，提醒远端的医护人员。这一方面使医护人

员不必每时每刻地关注看护对象的状态，而只需处理紧急情况；另一方面使病人在遇到突发情况时不必通过按钮等方式呼叫，因为那时病人可能失去了行动能力。

在汽车、火车和飞机等的行驶过程中，实时监测驾驶员的表情，对保障行驶安全有很大意义。可在车辆、飞机上安装表情识别设备，根据驾驶员的表情，提供智能服务。比如，当驾驶员表现出惊慌的神态时，系统自动播放节奏舒缓的音乐，放松人的心情；当驾驶员表现出疲劳、厌倦的神态时，系统会发出警报，并播放摇滚乐。

门禁系统常利用人脸、指纹、虹膜等信息识别人的身份，面部表情也可作为一种生物特征辅助识别。这是因为不同人在表达同一种情感时，其表情强度、产生部位、持续时间等特征存在差异。

肺部的空气经过声带进入口腔，最后由嘴辐射出声波，这就形成了语音。传统的语音识别技术包括说话内容识别和说话人识别，而语音信号中的情感信息往往被当作噪声过滤掉了。语音情感识别技术提取语音信号的情感特征，判断说话人的情感状态。由于民族文化和地域习惯的千差万别，不同地方的人们通过语音表达情感的方式存在较大差异。因此，需要依据本地区的实际情况，分析语音情感。东南大学是国内最早研究语音情感识别的机构；北京航空航天大学建立了汉语情感语音语料库。

情感识别技术应用于通信领域，将较大程度地节省带宽。发送端采集图像和语音，利用情感识别技术识别出表情和语音情感；网络上仅传输文字和情感状态；接收端建立虚拟的面部模型，模拟语音和表情。此外，在人机交互系统中，通过识别用户的语音情感，将使交互界面更加人性化和智能化；语音情感识别技术也可用于辅助有语言障碍的人表达情感。

人们通过面部表情、语音、肢体动作、文字等多种方式表达情感，因此情感识别本质上是一个多模态融合的问题。多模态情感识别采用模态融合技术，综合利用来自多个信息源的信息。常用的模态融合技术有三种类型，分别是特征级融合、模型级融合和决策级融合。特征级融合技术联合从多个模态中提取的情感特征，构建联合特征向量用于识别情感状态。由于各模态的时间尺度可能不同，这种方法将提高同步工作的难度；同时增加的特征维度也会降低特征级融合的时间性能。模型级融合技术为各模态建立统一的模型，既降低了对各信息源同步的要求，又利用了各模态之间的关联信息。大多数多模态情感识别采用决策级融合技术。决策级融合为各模态的情感分别建立模型，依据各模态独立识别的结果，得到模态融合后的情感状态。D-S 证据理论、人工神经网络等都是常用的决策级模态融合方法。

1.2 研究内容

本文设计算法实现了多模态融合的情感识别，分为情感特征提取和融合多模态的情感识别两个主要方面。情感特征提取包括表情特征提取和情感语音特征提取；融合多模态的情感识别利用了隐马尔可夫模型和人工神经网络中的多层感知器。

为从面部表情图像序列中提取表情特征，建立面部表情图像的主动外观模型，检测跟踪面部特征点的坐标，根据特征点位移提取面部动画参数，各帧表情图像的面部动画参数构成表情特征向量序列。根据训练集中面部特征点的坐标，建立反映面部表情图像中面部特征点形状变化的模型；同时利用面部区域的灰度外观，建立反映面部表情图像外观变化的模型；联合形状模型和灰度外观模型，构建统一的主动外观模型。利用主动外观模型，可以重建面部表情图像，进而检测出图像中的面部特征点。面部动画参数与面部特征点的运动相关，每个面部动画参数都是根据特征点在水平和垂直方向的位移得到的。

本文提取的情感语音特征包括短时平均能量、基音频率和共振峰等。由于语音信号具有短时平稳性，可对其作分帧处理，再分别对各帧信号作时域和频域。其中，采用时域分析可估算出各帧的短时平均能量和基音频率，采用频域分析可求得各帧的基音频率和共振峰。各帧情感语音信号的短时平均能量、基音频率和共振峰构成情感语音特征向量序列。

从训练样本中提取的表情特征和情感语音特征作为观察向量序列，用于训练各种表情和语音情感的隐马尔可夫模型。可以认为观察向量序列符合混合高斯分布，因此训练的 HMM 模型为连续 HMM 模型。为训练 HMM 模型，首先对模型参数作初始化，均一分割用于训练某种 HMM 模型的全部观察向量的状态序列，采用 K-Means 聚类算法，将各状态的观察向量分配给不同的高斯混合，根据状态包含的各高斯混合的权重、均值向量和协方差矩阵，计算各状态下观察向量的概率密度函数。对训练集中的观察向量序列的状态序列作 Viterbi 分割，并修正各状态下观察向量的概率密度函数，重复此过程，直至模型参数不再发生明显变化。利用训练出的 HMM 模型，采用前向算法，可以计算出观察向量序列关于各 HMM 模型的条件概率。这些条件概率将作为多层感知器的输入，用于训练多层感知器的权重矩阵。

多层感知器由输入层、输出层和一个或多个隐藏层组成，每层包含若干神经元，各神经元与前后两层的神经元直接相连；多层感知器的训练采用反向传播学习算法，包括传播和更新权重两个主要阶段。

从测试样本中提取表情特征和情感语音特征构成测试观察向量序列，利用前向算法计算观察向量序列关于各 HMM 模型的条件概率作为多层感知器的输入，

根据反向传播算法中传播阶段的算法计算多层感知器的输出，即可识别出测试样本的情感状态。

1.3 全文安排

本文分为六章。第一章为绪论，介绍课题背景和主要研究内容；第二章为文献综述，多模态情感识别的国内外研究现状；第三章为情感特征提取，介绍建立主动外观模型以提取面部动画参数等表情特征，以及对语音信号作时域和频域分析以提取情感语音特征；第四章为基于 HMM 模型和人工神经网络的多模态融合情感识别，介绍建立各种表情和语音情感的 HMM 模型，以及融合表情和情感语音模态的多层感知器；第五章为多模态融合情感识别实验，包括实验平台设计，实验结果，以及实验结论；第六章为课题总结与展望。

第二章 多模态情感识别研究现状

2.1 表情识别研究现状

社会心理学认为，面部表情是人际交流的重要模态，对情感表达的贡献最大。研究人员利用摄像机获取带有表情信息的人脸图像，从中提取表情特征，用于识别人的情感状态。很多研究人员通过对图像的整体或局部作变换得到表情特征，常用的图像变换包括傅里叶变换、离散余弦变换、小波变换、Gabor 变换、Haar 变换等。为准确提取面部表情特征，需要在图像中检测出人脸的位置，Ada-boost 算法是一种可用于人脸检测的快速算法。实验中常用的面部表情图像数据库包括 JAFFE 人脸库、Yale 人脸库、ORL 人脸库等。

崔景霞对人脸面部的不同器官作小波变换，分析了各器官情感表达的差异[1]。丁志起等人提出基于图像局部保留投影的表情识别方法，将面部表情图像的结构信息融入局部保留投影的目标函数，提取的表情特征更具判别性[2]。祝长生等人采用肤色模型在图像中定位人脸，利用基于 Canny 算子的主动外观模型定位面部特征点，通过曲线拟合方法计算表情特征，提取的特征向量用于表情识别具有较好的实时性和鲁棒性[3]。张杰对面部表情图像作 Gabor 小波变换，建立表情的局域弹性模板，对待测图像与表情模板作非刚性匹配，采用 k-近邻法识别表情[4]。朱明旱等人提出基于图像重建的表情识别方法。从面部表情图像中提取表情流形，建立流形与图像的映射；确定图像在流形空间中的坐标；利用图像在表情路径的投影，重建并识别表情[5]。郑秋梅等人提出基于改进的线性判别分析的表情识别方法。改进的线性判别分析从水平和垂直两个维度对图像作线性判别分析，采用主成分分析法降低特征向量的维数[6]。刘松等人提出融合局部特征和整体特征的面部表情算法。采用 Fisher 线性判别分析从面部表情图像中提取整体特征；利用主成分分析和 Fisher 线性判别分析提取局部特征；以人工神经网络作为分类器识别表情。算法在 Yale 和 JAFFE 人脸库上的实验具有较高的识别精度[7]。张建明等人设计了从带遮挡的人脸图像中提取表情特征的方法，采用中心线检测算法判断面部表情的二值图像是否存在遮挡，通过对称变换修复被遮挡的表情图像。利用修复后的表情图像，可以获得更高的表情识别率[8]。周川等人认为人的身份和表情是决定面部表情图像的主要因素，提出因素分解模型，在身份子空间和表情子空间中利用余弦距离识别表情状态[9]。

Bassili J 的实验证实，人们识别图像序列中面部表情的能力，相比于识别静态图像中的面部表情更高[10]。因此，研究人员提出利用带有表情的图像序列识别情感状态的方法。实验中常用的包含面部表情图像序列的数据库有 CMU PIE 人

脸库和 Cohn-Kanade 人脸表情库等。

表情的产生源于面部肌肉组织运动，Ekman 和 Friesen 建立面部运动编码体系(Facial Action Coding System, FACS), 为每一块面部肌肉的收缩运动编码。面部运动编码体系的测量单元是动作单元(Action Unit, AU), 而不是面部肌肉[11]。Tian Y 等人根据面部运动编码体系，利用面部表情图像序列，计算面部各动作单元的肌肉收缩强度，用于识别表情状态[12]。

很多方法通过检测和跟踪面部特征点在人脸图像序列中的坐标位置，计算特征点的位移作为表情特征。Pardas M 等人提出面部定义参数(Facial Definition Parameters, FDP)和面部动画参数(Facial Animation Parameters, FAP). 面部定义参数记录了三维人脸图像中的特征点，这些特征点来自眉、眼、鼻、唇、耳、面颊、头顶等诸多部位，面部定义参数描述了这些特征点的相对位置；面部动画参数表达了面部特征点的运动，由面部定义参数中的面部特征点在水平、垂直或前后方向的位移计算得到，可直接作为表情特征；面部定义参数和面部动画参数皆与 MPEG4 标准兼容[13-14]。为准确计算面部动画参数等与面部特征点运动相关的表情特征，需要检测和跟踪面部特征点的坐标位置。常用的刚体特征点检测算法包括轮廓线算法、主动形状模型和主动外观模型等。Pardas M 等人提取面部动画参数作为表情特征，利用隐马尔可夫模型识别图像序列中的表情[15]。

余棉水等人计算面部图像序列的光流作为表情特征，采用人工神经网络作为分类器识别情感[16]。杨国亮等人将 Hessian 矩阵 Lucas-Kanade 光流法，提高了光流场的计算精度，并结合隐马尔可夫模型识别图像序列中的表情[17]。丁志起等人计算图像序列相邻帧之间表情子区域的差图像，对差图像作 Gabor 小波变换作为表情特征，并利用下采样降维法减少特征向量的维数，采用支持向量机作为分类器识别面部表情图像序列中的情感状态[18]。

2.2 语音情感识别相关研究

在人们对话交流的过程中，语音中的情感对情感表达也有相当大的贡献。研究人员从语音信号中提取情感语音特征，用于识别情感状态。常用的情感语音特征包括振幅、能量、语音持续时间、基音频率、共振峰、梅尔频率倒谱系数、梅尔频域子带能量、线性预测倒谱系数等等。通过对语音信号作时域、频域以及倒谱域分析，可以提取这些特征。大部分方法从语音信号的每一帧提取特征，构成特征序列用于识别情感；另外一些方法计算各帧特征的统计量，如均值、方差等，并计算各统计量的各阶导数，一并作为情感特征。

赵力等人选取语句发音持续时间以及与基音频率、能量和共振峰有关的特征量构成情感语音特征向量，其中，基音频率相关的特征量包括平均基音频率、最

大基音频率以及基音频率一阶导数的平均值，能量相关的特征量包括平均能量和能量的动态范围，共振峰相关的特征量包括共振峰的平均值、共振峰一阶导数的平均值、共振峰峰值点回归直线斜率的平均值以及共振峰峰值的平均值。计算待测样本与训练集中各情感，类别样本均值的马氏距离，距离最近的情感类别即为识别结果[19, 20]。Lin Y 等人从语音信号中提取基音频率、共振峰(F1-F4)、梅尔频域倒谱系数(MFCC1，MFCC2)、梅尔频域子带能量(MBE1-MBE5)，并计算它们的一阶导数和二阶导数，组成 39 维的特征向量；采用序列前向选择算法选取其中的五种特征，构成降维后的特征向量；实验分别采用隐马尔可夫模型和支持向量机作为情感分类器，并对结果作了比较[21]。屠彬彬等人提取改进的梅尔频域倒谱系数作为语音特征。利用经验模态分解得到语音信号的固有模态函数分量，通过梅尔滤波器选取其对数能量，作离散余弦反变换得到改进的梅尔频域倒谱系数。实验结果表明利用改进的 MFCC 特征识别情感具有较高的识别率[22]。叶吉祥等人提取 hurst 指数和多重分型谱参数作为情感语音特征，采用支持向量机对情感作分类[23]。康燕等人比较了在各种噪声环境下，ZCPA 参数与其它情感语音特征用于情感识别的识别率，分类器采用连续的隐马尔可夫模型。结果表明，ZCPA 参数是一种可以用于语音情感识别的特征[24]。余华等人将混合蛙跳算法引入人工神经网络的训练过程，用于对语音情感状态作分类，同时比较了混合蛙跳算法神经网络与 BP 神经网络、RBF 神经网络的语音情感识别性能[25]。张石清等人基于改进的监督流形学习算法的语音情感识别方法，改进后的算法具有最优泛化能力，同时增强了低位嵌入数据的判别力，用于语音情感识别具有得到较高的识别率[26]。

2.3 模态融合相关工作

心理学研究表明，人们主要依赖面部表情和语音判断一个人的情感状态，情感识别本质上是多模态问题。模态融合技术主要有三类，分别是特征级融合、模型级融合和决策级融合。

特征级融合分别从多种模态提取特征，构建用于识别情感的联合特征向量，对各模态有较高的同步要求。黄程韦等人提出融合语音信号和心电信号的情感识别，提取情感语音特征和心率变异性特征作为情感特征，实验表明，基于语音信号和心电信号的多模态分类器比单模态分类器具有更高的情感识别精度[27]。

模型级融合既没有对多模态信息流的同步要求，又利用了各模态信息流之间的关联信息。Zeng Z 等人提出多流融合隐马尔可夫模型用于情感识别。多流融合隐马尔可夫模型是双流融合隐马尔可夫模型的泛化，是一种通用的模型级模态融合手段。多流融合隐马尔可夫模型的优点如下。来自每个模态的特征都能用一个组件隐马尔可夫模型建模，根据最大熵准则和最大互信息评价标准，这个组件

隐马尔可夫模型与其它组件之间有最优的连接；不同组件隐马尔可夫模型的状态转移不一定同时发生；如果一个组件隐马尔可夫模型损坏了，其它组件仍能正常工作；相比于其它基于隐马尔可夫模型的融合方法，多流融合隐马尔可夫模型在复杂度和性能之间有更好的平衡[28-30]。

大多数多模态融合情感识别采用决策级融合。决策级融合为音频情感和视频情感单独建模，然后联合这些单模态的识别结果。Hoch S 等人采用加权求和的方法，在决策级融合表情和情感语音模态[31]。Go H J 等人通过线性叠加的方式，合并了表情和语音情感各自的识别结果[32]。

2.4 本章小结

研究人员在表情特征提取、语音特征提取以及情感识别等方面做了大量的工作。本文提出多模态融合的情感识别算法，在表情模态提取面部动画参数特征，在语音模态提取能量、基音频率和共振峰等特征，并基于隐马尔可夫模型和人工神经网络设计了模态融合算法。

第三章 表情和语音特征提取

特征提取是多模态情感识别的重要步骤，表情和语音特征的区分性直接决定了情感识别的准确程度，特征提取算法的时间复杂度对情感识别的实时性有很大影响。

3.1 建立面部表情图像的主动外观模型

本文提出基于主动外观模型提取表情特征的方法。基于主动外观模型的表情特征提取步骤如下。首先，利用一系列带有特征点标注的面部图像，训练出主动外观模型；其次，利用主动外观模型，在人脸表情图像序列中检测面部特征点的坐标位置；再次，根据图像序列中的特征点在水平和垂直方向的位移，计算面部动画参数作为表情特征向量。

3.1.1 特征点检测方法

主动外观模型是 T. F. Cootes 提出的一种为弹性物体图像建模的方法[33-35]，它由主动形状模型发展而来。Cootes 认为，同一类弹性物体在不同的情况下，可能表现为不同的形状。例如，在医疗图像中，同一种人体器官的形状是因人而异的，且即使同一个人在不同时期拍摄的器官，其形状也千差万别；又如，同一类工业配件由于型号的不同，其形状和外观都各有不同。总而言之，允许弹性物体的形状在一定程度内发生改变。由此，Cootes 提出为图像中形状可变的物体建立弹性模型，该模型总结了同一类型物体形状变化的规律，允许物体的形状发生较大的变化。

主动形状模型即是满足上述要求的这样一种模型。主动形状模型训练集中的每一个样本是一系列代表图像中弹性物体形状的点，这些点通常处于物体的灰度边缘；不同样本中的训练点是对应的，即它们在物体上的相对位置是一致的；训练点的坐标位置的标注是手工完成的。不同样本中对应的训练点需要按照最小化对应点距离的原则对齐，对齐的过程中对物体形状进行放缩、旋转和平移，并利用最小二乘法实现对应点距离的加权平方和最小。计算对齐后训练样本的统计量，即得到主动形状模型，这一过程利用了主成分分析的方法，由对齐后的训练点坐标构建样本矩阵，计算样本矩阵的协方差矩阵，协方差矩阵的特征向量构成了训练集的特征空间。

利用训练得到的主动形状模型，可检测出一幅图像中弹性物体的特征点坐标位置。首先采用某种方法初始化特征点的坐标，如 Cootes 提出的基于遗传算法

的初始化方法；然后依据特征点附近灰度与训练集对应点附近平均灰度的匹配程度，迭代地调整特征点坐标，直至检测出的物体形状不再发生明显变化[36-38]。

3.1.2 建立形状模型

主动外观模型在标准化形状的框架下，联合了形状变化的模型和外观变化的模型。训练集中的每个样本都是一幅带有标注的图像，标注文件标出了图像中物体特征点的坐标位置。比如，为面部表情图像建立主动外观模型，训练样本通常为不同光照强度、姿态、表情的人脸图像，面部特征点的坐标位置被记录在标注文件中。

建立面部表情图像的训练集后，利用标注文件中对应点的坐标位置，为面部特征点组成的形状建模，即建立一个反映训练样本形状变化的统计模型。

为比较不同训练样本中特征点组成的形状，需要将训练集中所有样本特征点组成的形状对齐，使对齐后各样本对应特征点距离的加权平方和最小。这一对齐的过程通过形状的放缩、旋转和平移变换实现，并采用了最小二乘法。

考虑对齐两个训练样本的过程。设 p_i 表示第 i 个训练样本的 n 个特征点横、纵坐标构成的向量，$p_i=(x_{i0},y_{i0},x_{i1},y_{i1},...,x_{ik},y_{ik},...,x_{i(n-1)},y_{i(n-1)})^T$，$(i=1,2,\cdots,m)$.

给定 p_i 和 p_j，选择合适的放大倍数 s_j、旋转角度 φ_j 以及平移量 (t_{xj},t_{yj})，将 p_j 映射到 $L(s_j,\varphi_j)[p_j]+t_j$，使得 p_i 与 $L(s_j,\varphi_j)[p_j]+t_j$ 距离的加权平方和

$$D_j=(p_i-L(s_j,\varphi_j)[p_j]-t_j)^T W(p_i-L(s_j,\varphi_j)[p_j]-t_j)$$

最小。其中，

$$L(s,\varphi)\begin{bmatrix}x_{jk}\\y_{jk}\end{bmatrix}=\begin{bmatrix}(s\cos\varphi)x_{jk}-(s\sin\varphi)y_{jk}\\(s\sin\varphi)x_{jk}+(s\cos\varphi)y_{jk}\end{bmatrix},$$

$$t_j=(t_{xj},t_{yj},...t_{xj},t_{yj})^T,$$

$W=\begin{bmatrix}w_0 & & & \\ & w_1 & & \\ & & \ddots & \\ & & & w_{n-1}\end{bmatrix}$ 为权重对角阵，$w_k=(\sum_{m=0}^{n-1}V_{R_{km}})^{-1}$，$(k=0,1,..n-1)$，而 r_{km} 为某训练样本中第 k 个特征点和第 m 个特征点之间的距离，$V_{r_{km}}$ 是全部训练样本的 r_{km} 的方差。若第 k 个特征点的方差和 $\sum_{m=0}^{n-1}V_{r_{km}}$ 较大，则说明该特征点的相对

位置变化较大，分配给它的权重 w_k 就比较小；反之，若 $\sum_{m=0}^{n-1} V_{r_{km}}$ 较小，则说明该特征点的相对位置变化较小，分配给它的权重 w_k 就比较大。

为求出使距离的加权平方和 D_j 最小的放大倍数 s_j、旋转角度 φ_j 以及平移量 (t_{xj}, t_{yj})，采用最小二乘法，得到线性方程组

$$\begin{bmatrix} X_2 & -Y_2 & W & 0 \\ Y_2 & X_2 & 0 & W \\ A & 0 & X_2 & Y_2 \\ 0 & A & -Y_2 & X_2 \end{bmatrix} \begin{bmatrix} u_j \\ v_j \\ t_{xj} \\ t_{yj} \end{bmatrix} = \begin{bmatrix} X_1 \\ Y_1 \\ B \\ C \end{bmatrix}.$$

其中，

$u_j = s_j \cos\varphi_j$，$v_j = s_j \sin\varphi_j$，

$$X_1 = \sum_{k=0}^{n-1} w_k x_{ik}\ ,\ X_2 = \sum_{k=0}^{n-1} w_k x_{jk}\ ,\ Y_1 = \sum_{k=0}^{n-1} w_k y_{ik}\ ,\ Y_2 = \sum_{k=0}^{n-1} w_k y_{jk}\ ,\ W = \sum_{k=0}^{n-1} w_k\ ,$$

$$A = \sum_{k=0}^{n-1} w_k (x_{jk}^2 + y_{jk}^2),\ B = \sum_{k=0}^{n-1} w_k (x_{ik}x_{jk} + y_{ik}y_{jk}),\ C = \sum_{k=0}^{n-1} w_k (y_{ik}x_{jk} - x_{ik}y_{jk}).$$

求解该线性方程组，可得到 u_j，v_j，t_{xj}，t_{yj}，进而求出 s_j 和 φ_j.

设训练集中的样本特征点构成的向量有 p_1，p_2，…，p_m，采用如下算法将训练集中的全部样本对齐。

1. 将样本 p_2，…，p_m 对齐到样本 p_1.

2. 计算对齐后的 m 个样本的平均形状 $\bar{p} = \frac{1}{m}\sum_{i=1}^{m} p_i$.

3. 标准化平均形状 $\bar{p}$.

3.1 对 $\bar{p}$ 作放缩变换，使其某两个特征点的距离为一固定值。

3.2 对 $\bar{p}$ 作旋转变换，使其某两个特征点的连线的方向为一固定的方向。

3.3 对 $\bar{p}$ 作平移变换，使其所有特征点的几何中心为一固定的位置。

4. 将 $p_1, p_2, \ldots, p_m$ 对齐到 $\bar{p}$.

5. 重复 2-5，直至 $\bar{p}$ 不再发生明显变化。

对平均形状 $\bar{p}$ 作标准化是必要的，因为若不进行标准化，则算法的 4m 个变量只有 4(m-1)个约束，这将导致算法陷入无限次的迭代。

图 3-1 为以 IMM 人脸数据库为训练集，将训练样本的特征点按迭代算法对齐后得到的平均形状 $\bar{p}$.

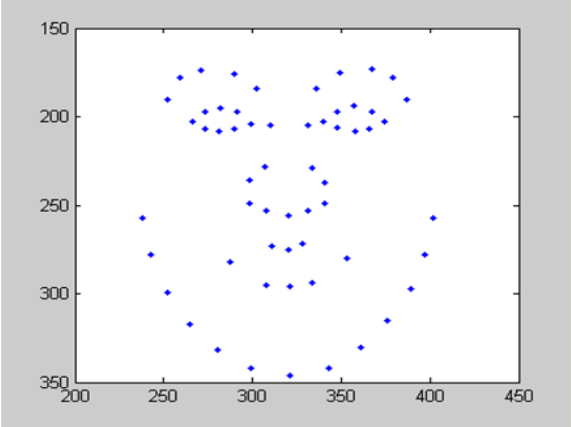

图 3-1 平均形状

在对齐后的训练样本 p_1，p_2，…，p_m 上作主成分分析，可得到反映训练样本形状变化的统计模型。

求出对齐后的训练样本 p_1，p_2，…，p_m 与平均形状 $\bar{p}$ 之间的差值 $dp_i = p_i - \bar{p}$，$(i = 1,2,\cdots,m)$. 计算训练样本的协方差矩阵 $\Sigma_s = \frac{1}{m}\sum_{i=1}^{m} dp_i\, dp_i^T$. 由线性方程组 $\Sigma_s u_{sk} = \lambda_k u_{sk}$，$u_{sk}^T u_{sk} = 1$，$(k = 1,2,\cdots,2n)$，可得协方差矩阵 Σ_s 的特征值 λ_k $(\lambda_{k-1} \geq \lambda_k, k = 2,3,\cdots,n)$ 和特征向量 u_{sk}.

某样本的形状可由下式近似计算出，

$$p = \bar{p} + U_s b_s, \tag{3.1}$$

其中，$U_s = (u_{s1}, u_{s2}, \cdots, u_{st})$ 由 Σ_s 的前 t 个特征向量组成，$b_s = (b_{s1}, b_{s2}, \cdots, b_{st})$ 是表示权重的向量。通过改变参数向量 b_s，可以得到新样本，b_s 的取值受到如下约束的限制，

$$-3\sqrt{\lambda_k} \leq b_{sk} \leq 3\sqrt{\lambda_k}.$$

图 3-2 分别为b_1取$[-3\sqrt{\lambda_1},3\sqrt{\lambda_1}]$范围内的 9 个不同的值，$b_2,\cdots,b_t$都取 0 时，得到的样本特征点组成的形状，反映了第一主成分对形状的影响。由图可知，对面部表情图像中面部特征点的形状作主成分分析，第一主成分代表人脸旋转角度。

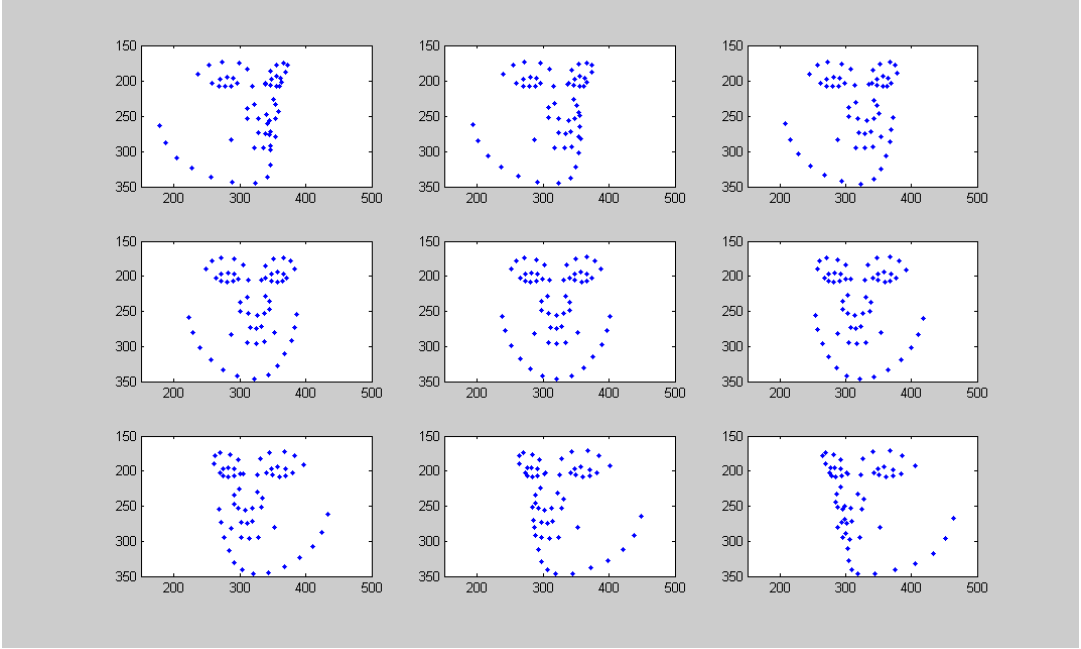

图 3-2 第一主成分对形状的影响

图 3-3 分别为b_2取$[-3\sqrt{\lambda_2},3\sqrt{\lambda_2}]$范围内的 9 个不同的值，$b_1,b_3,\cdots,b_t$都取 0 时，得到的样本特征点组成的形状，反映了第二主成分对形状的影响。由图可知，随着b_2的增加，下巴由尖变圆，鼻子宽度增加，眼睛由大变小，眉毛弧度增大。

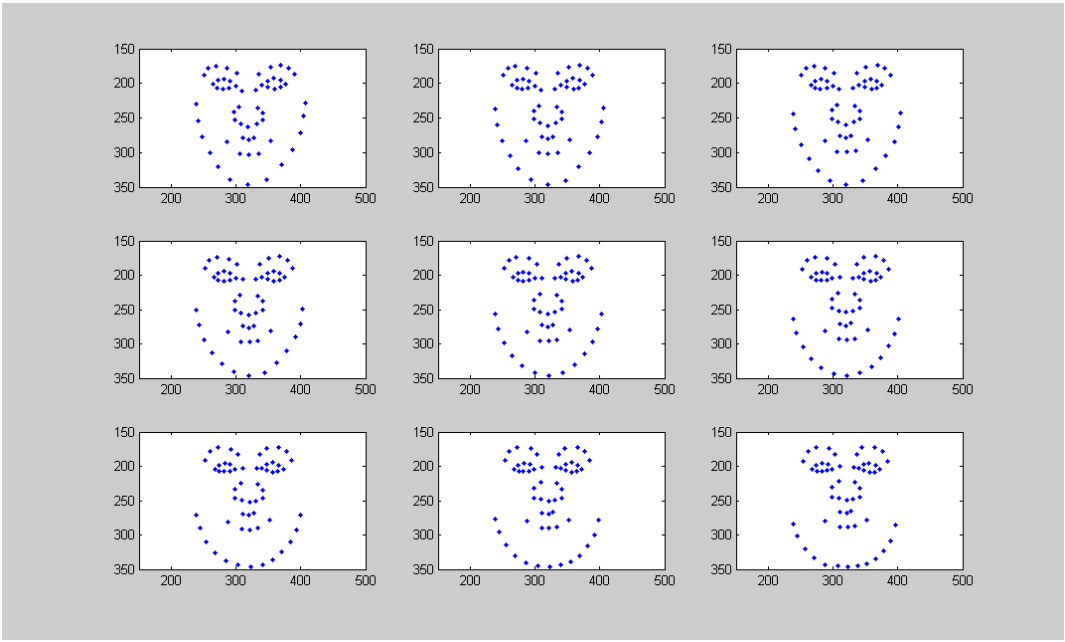

图 3-3 第二主成分对形状的影响

图 3-4 分别为b_3取$[-3\sqrt{\lambda_3},3\sqrt{\lambda_3}]$范围内的 9 个不同的值，$b_1,b_2,b_4,\cdots,b_t$都取 0 时，得到的样本特征点组成的形状，反映了第三主成分对形状的影响。由图可知，随着b_2的增加，下巴由尖变圆，鼻子宽度增加，眼睛由小变大，眉毛弧度变小。

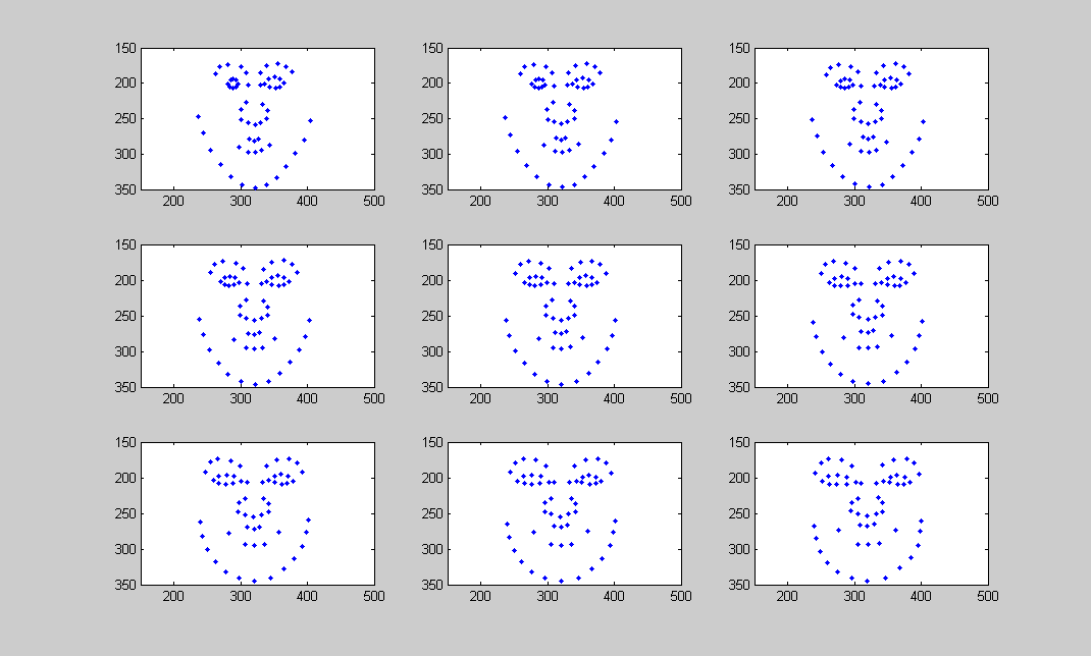

图 3-4 第三主成分对形状的影响

至此，建立了反映样本形状变化的统计模型。

3.1.3 建立外观模型

下面对面部区域的灰度外观建模，即建立一个反映训练样本外观变化的统计模型。

为比较不同训练样本中面部区域的灰度外观，对每幅面部表情图像中的像素按三角形法作变换，使样本中的特征点与平均形状中的特征点对齐。

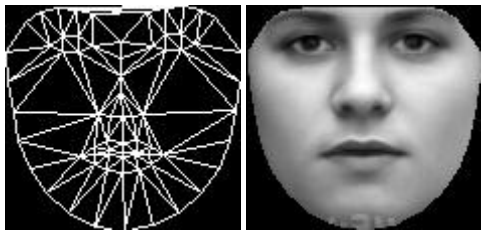

图 3-5 主动外观模型

设训练样本按三角形法对齐到平均形状后，面部区域各像素的灰度构成向量g_i，$(i=1,2,\cdots,m)$. 在对齐后的训练样本 g_1，g_2，…，g_m上作主成分分析，可得到反映训练样本形状变化的统计模型。图 3-5 中的左图为按照三角形法分割的面部

区域，右图为以 CK2010 库为训练集，按三角形法将样本中的特征点对齐后得到的面部区域的平均灰度。

求出对齐后的训练样本 g_1，g_2，…，g_m 与平均灰度 $\overline{g}$ 之间的差值 $dg_i = g_i - \overline{g}$. 计算训练样本的协方差矩阵 $\Sigma_g = \frac{1}{m}\sum_{i=1}^{m} dg_i\, dg_i^T$. 由 $\Sigma_g u_{gk} = \lambda_k u_{gk}$，$u_{gk}^T u_{gk} = 1$，$(k = 1,2,\cdots,2n)$，可得协方差矩阵 Σ_g 的特征值 λ_k $(\lambda_{k-1} \geq \lambda_k, k = 2,3,\cdots,2n)$ 和特征向量 u_{gk}.

某样本的灰度可由下式近似计算出，

$$g = \overline{g} + U_g b_g, \tag{3.2}$$

其中，$U_g = (u_{g1}, u_{g2}, \cdots, u_{gt})$ 由 Σ_g 的前 t 个特征向量组成，$b_g = (b_{g1}, b_{g2}, \cdots, b_{gt})$ 是表示权重的向量。通过改变参数向量 b_g，可以得到新样本，b_g 的取值受到如下约束的限制，

$$-3\sqrt{\lambda_k} \leq b_{gk} \leq 3\sqrt{\lambda_k}.$$

这样，反映样本灰度外观变化的统计模型就建立起来了。

3.1.4 联合形状和外观建立模型

由于面部表情图像中的特征点组成的形状与面部区域的灰度外观之间有一定的关联，因此，联合上面建立的形状和外观模型，构建统一的主动外观模型。

由(3.1), (3.2), 得

$$b_{si} = U_s^T (p_i - \overline{p_i}),$$

$$b_{gi} = U_g^T (g_i - \overline{g_i}).$$

设 $b_i = \begin{bmatrix} b_{si} \\ b_{gi} \end{bmatrix}$，$(i = 1,2,\cdots,m)$，对 b_1，b_2，…，b_m 作主成分分析，可得

$$b = Vc,$$

其中，$V = \begin{bmatrix} V_s \\ V_g \end{bmatrix}$. 于是，(3.1), (3.2)可改写为

$$p = \overline{p} + U_s V_s c, \tag{3.3}$$

$$g = \bar{g} + U_g V_g c. \tag{3.4}$$

由上面两式可知，改变参数向量 c，可以得到新样本，这些新样本具有不同的特征点组成的形状，以及不同的面部区域的灰度外观。通过调整参数向量到某个合适的值，我们可以实现面部表情图像的重建，以及检测图像中面部特征点的坐标位置。

3.2 面部特征点检测

利用主动外观模型，可以重建面部表情图像，进而检测出图像中的面部特征点。基本思路如下。为主动外观模型设定参数向量的初始值，得到初始的合成图像；迭代地调整参数向量的值，使得合成图像与原面部表情图像尽可能相同；利用参数向量，计算原图像中面部特征点的坐标位置。

重建面部表情图像的目标是使合成图像与原图像尽量接近。设原图像面部区域各像素灰度构成的向量为 g_{origin}，合成图像面部区域各像素灰度构成的向量为 $g_{reconstruct}$，则原面部表情图像与合成图像面部区域各像素灰度的差值向量为 $\Delta g = g_{origin} - g_{reconstruct}$. 为使重建图像与原图像尽可能接近，则需使 $|\Delta g|$ 的值尽可能地小。

设主动外观模型的参数向量为 c_0，参数向量的增量为 Δc，则得到合成图像新的参数向量为 $c = c_0 + \Delta c$. 由公式(3)，(4)，得到合成图像的特征点坐标构成的向量 p，以及面部区域灰度构成的向量 g. 取原图像面部区域的各像素灰度构成向量 g_o. 设 $\Delta g = g_o - g$，则根据不同的 Δc，可以计算出不同的 Δg.

假设 Δc 与 Δg 之间存在线性关系 $\Delta c = A\Delta g$，通过线性回归分析，可以计算出系数 A 的值。

根据初始参数向量 c_0 和系数 A，采用如下的迭代算法，可以得到与原面部表情图像相近的合成图像。

1. 由初始参数向量 c_0，计算误差向量 $\Delta g_0 = g_o - g$.

2. 计算参数向量增量 $\Delta c = A\Delta g_0$.

3. 设新的参数向量 $c_1 = c_0 - k\Delta c$, $k = 1$, 由 c_1 按第 1 步中的公式计算新的误差向量 Δg_1. 若 $|\Delta g_1| < |\Delta g_0|$, 则接受新的参数向量 c_1；否则，设 $k = 1.5, 0.5, 0.25, \ldots$, 重新计算 c_1.

4. 重复 1-3，直至 $|\Delta g|$ 不再有明显变化。

采用上述迭代算法，得到合成图像的参数向量 c；根据公式(3.3), (3.4), 可以计算出合成表情图像面部区域的像素灰度，以及面部特征点的坐标位置。

图 3-6 面部图像重建和特征点检测

选取 CK2010 库中若干幅面部表情图像构成训练集，构建主动外观模型。对测试集中的面部表情图像，利用主动外观模型，采用上述迭代算法作图像重建和面部特征点检测，结果如图 3-6 所示。其中，各行的左图为事先标定的面部特征

点的坐标位置，中图为利用主动外观模型对图像作重建的结果，右图为采用迭代算法检测出的面部特征点的坐标位置。

实验表明，训练图像的选取对图像重建和特征点检测的效果有较大影响。若训练集只包含表情平静的图像，则对带有表情的图像作重建时准确性较低；若训练集只包含带有轻微表情的图像，则重建表情强烈的图像时也会有较大的误差。因此，本文选取多名志愿者带有不同表情的图像构造训练集，在检测面部特征点时取得了较好的效果。

3.3 面部动画参数提取

面部动画参数是一套与 ISO MPEG-4 标准兼容的参数集，与面部肌肉运动相关，能够表达基本的面部动作，比如皱眉、眨眼、张口、闭口等，因而可以表达大多数面部表情。

面部动画参数与面部特征点的运动相关，68 个面部动画参数的每一个都是根据特征点在水平和垂直方向的位移计算得到。比如，FAP4(lower_t_midlip)是由内唇上点垂直位移算得，FAP19(close_t_l_eyelid)是由左眼皮上点垂直位移得到，FAP31(raise_l_i_eyebrow) 是由左眉内点垂直位移计算得出，而FAP51(lower_t_midlip_o)则是由外唇上点垂直位移算出。因而，可以利用主动外观模型在表情图像序列中检测出面部特征点，进而根据特征点坐标位置的变化计算面部动画参数。表 3-1、表 3-2、表 3-3 和表 3-4 分别记录了与内唇点、眉点、外唇点、眼睑点运动相关的面部动画参数的编号、名称和计算方法。

利用面部动画参数单元可以将面部动画参数标准化。面部动画参数单元使得所有面部模型上的面部动画参数具有一致的评价。面部动画参数单元 ENS、ES、IRISD、MNS 以及 MW 的定义如图 3-7 所示。

$$ENS = ENS0/1024$$

$$ES = ES0/1024$$

$$IRISD = IRISD0/1024$$

$$MNS = MNS0/1024$$

$$MW = MW0/1024$$

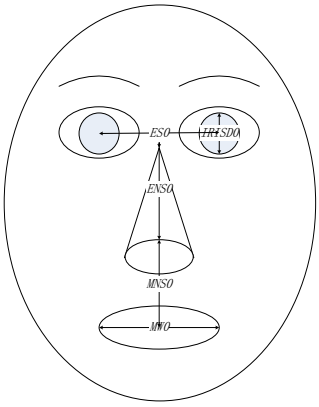

图 3-7 面部动画参数单元

表 3-1 记录了与内唇点运动相关的面部动画参数及其面部动画参数单元(FAPU).

表 3-1 与内唇点运动相关的面部动画参数

编号	名称	描述	FAPU	方向
4	lower_t_midlip	内唇上点垂直位移	MNS	垂直向下
5	raise_b_midlip	内唇下点垂直位移	MNS	垂直向上
6	stretch_l_cornerlip	内唇左点水平位移	MW	水平向左
7	stretch_r_cornerlip	内唇右点水平位移	MW	水平向右
8	lower_t_lip_lm	内唇左上点垂直位移	MNS	垂直向下
9	lower_t_lip_rm	内唇右上点垂直位移	MNS	垂直向下
10	raise_b_lip_lm	内唇左下点垂直位移	MNS	垂直向上
11	raise_b_lip_rm	内唇右下点垂直位移	MNS	垂直向上
12	raise_l_cornerlip	内唇左点垂直位移	MNS	垂直向上
13	raise_r_cornerlip	内唇右点垂直位移	MNS	垂直向上

表 3-2 描述了与眉点运动相关的面部动画参数。

表 3-2 与眉点运动相关的面部动画参数

编号	名称	描述	FAPU	方向
31	raise_l_i_eyebrow	左眉内点垂直位移	ENS	垂直向上
32	raise_r_i_eyebrow	右眉内点垂直位移	ENS	垂直向上
33	raise_l_m_eyebrow	左眉中点垂直位移	ENS	垂直向上
34	raise_r_m_eyebrow	右眉中点垂直位移	ENS	垂直向上
35	raise_l_o_eyebrow	左眉外点垂直位移	ENS	垂直向上
36	raise_r_o_eyebrow	右眉外点垂直位移	ENS	垂直向上
37	squeeze_l_eyebrow	左眉水平位移	ES	水平向右
38	squeeze_r_eyebrow	右眉水平位移	ES	水平向左

表 3-3 描述了与外唇点运动相关的面部动画参数。

表 3-3 与外唇点运动相关的面部动画参数

编号	名称	描述	FAPU	方向
51	lower_t_midlip_o	外唇上点垂直位移	MNS	垂直向下
52	raise_t_midlip_o	外唇下点垂直位移	MNS	垂直向上
53	stretch_l_cornerlip_o	外唇左点水平位移	MW	水平向左
54	stretch_r_cornerlip_o	外唇右点水平位移	MW	水平向右
55	lower_t_lip_lm_o	外唇左上点垂直位移	MNS	垂直向下
56	lower_t_lip_rm_o	外唇右上点垂直位移	MNS	垂直向下
57	raise_t_lip_lm_o	外唇左下点垂直位移	MNS	垂直向上
58	raise_t_lip_rm_o	外唇右下点垂直位移	MNS	垂直向上
59	raise_l_cornerlip_o	外唇左点垂直位移	MNS	垂直向上
60	raise_r_cornerlip_o	外唇右点垂直位移	MNS	垂直向上

表 3-4 描述了与眼睑点运动相关的面部动画参数。

表 3-4 与眼睑点运动相关的面部动画参数

编号	名称	描述	FAPU	方向
19	close_t_l_eyelid	左眼皮上点垂直位移	ENS	垂直向下
20	close_t_r_eyelid	右眼皮上点垂直位移	ENS	垂直向下
21	close_b_l_eyelid	左眼皮下点垂直位移	ENS	垂直向上
22	close_b_r_eyelid	右眼皮下点垂直位移	ENS	垂直向上

由于面部动画参数能够较好地区分不同类型的面部表情，而且计算过程简单、可靠，因此，可以提取面部动画参数作为表情特征，用于识别面部表情。

3.4 语音的情感特征

在人们对话交流的过程中，语音对情感表达的贡献占有相当大的比重。通过说话声音音调的高低、响度的强弱、音色的变化，可以判断出说话人当时的情感状态和心理波动。相同的说话内容，可以采用不同的说话方式表达，进而产生了多种语音情感。常见的语音情感状态包括高兴、愤怒、悲伤、害怕、惊讶、厌恶等。为准确识别出一段语音信号中说话人的情感状态，需要提取具有区分性的语音情感特征。本文提取的语音情感特征包括能量、基音频率和共振峰等。

3.4.1 语音信号的产生机制

人的发声器官分为气源、声门和声道三部分。其中，气源指肺和气管，声门指喉和声带，声道指咽腔、口腔和鼻腔。声门以内，称为声门子系统，负责产生激励振动；声门以外，包括声道系统和辐射系统。肺部的空气进入喉部，经过声带，产生声波；声波进入声道，最后由嘴辐射出，这就形成了语音。

由于发声方式的不同，人们发出的语音分为很多种。当声带收缩时，肺部的空气经过声带会引起声带的振动，这时发出的声音称为浊音；当声带舒张时，肺部的空气经过声带不会引起其振动，这时发出的声音称为清音。气流经过声道时遇到收紧点，产生的清音叫做摩擦音；气流经过声道时被暂时阻止而后又突然释放，产生的清音叫做爆破音。浊音由元音和浊辅音组成，清音由清辅音组成。浊音振幅较强，且具有准周期性；清音振幅较弱，波形类似高斯噪声。

声带振动是一种机械振动。机械振动在介质中传播，产生机械波；声带振动在空气中传播，产生声波。根据简谐运动的描述公式

$$x = A\sin(\omega t + \varphi),$$

机械振动可以由三种物理量描述，分别是振幅、频率和相位。这些物理量对应着人耳对声音的感知。其中，振幅对应人耳听到声音的响度，即人们常说的“分贝数”；频率对应人耳感知声音的音调，如低音 dol, 高音 re 等等，人耳能听到的振

动的频率范围是 20Hz-20000Hz.

浊音是由声带振动产生的，声带开启和闭合一次的时间间隔称为基音周期，基音周期的倒数即基音频率。基音频率的决定因素包括声带的尺寸、厚度、松紧程度及声带两侧的空气压强差。浊音的基音周期范围是 80Hz-500Hz.

当物体作受迫振动时，若驱动频率等于固有频率，物体会以最大振幅振动，这种现象称为共振。若声带振动频率等于声道的固有频率，则会引起声道的共振。声道的固有频率称为共振频率，或共振峰。声道具有很多共振峰，且每一时刻的共振峰由当时的声道形状决定。

3.4.2 语音信号的短时平稳性

由于发声器官的惯性运动，可以认为在 10-30ms 的时间里，语音信号近似不变，即语音信号具有短时平稳性。这样，可以把语音信号分为一系列分析帧进行处理，一般每秒的帧数约为 33-100 帧。

分帧时可以采用连续分段的方法，也可采用交叠分段的方法，这是为了使帧与帧之间平滑过渡，保持其连续性。交叠部分称为帧移，帧移与帧长比值的取范围是(0,1/2].

语音信号的分帧采用可移动窗口进行加权的方法实现，不同的窗函数形状会影响分帧后提取的短时特征。常用的窗有两种，分别是矩形窗和汉明窗。矩形窗的窗函数为 $\omega(n)=\begin{cases}1, 0\le n\le N-1\\ 0, \text{其他}\end{cases}$，矩形窗窗内所有采样点权重相同；汉明窗的窗函数为 $\omega(n)=\begin{cases}0.54-0.46\cos[2\pi n/(N-1)], 0\le n\le N-\\ 0, \text{其他}\end{cases}$. 与矩形窗相比，汉明窗能够保留更多的波形细节，因此更为常用。

窗口长度对能否反映语音信号的幅度变化起决定作用。若窗口长度过小，则窗函数等效于很宽的低通滤波器，不能得到比较平滑的短时特征；若窗口长度过大，比如达到若干个基音周期，则窗函数等效于很窄的低通滤波器，不能得到短时特征变化的细节。通常，一个语音帧应含有 1-7 个基音周期，根据语音基音周期的一般范围，在采样频率为 10kHz 的情况下，窗口长度取 100-200 个采样点。

本文提取的情感语音特征，如短时平均能量、基音频率和共振峰等，都属于短时特征，即它们都是在每一帧语音信号上分别计算得到的。

3.5 情感语音的时域分析

语音信号最直观的表示是它的时域波形。图 3-8 为 Berlin 情感语音数据库中

一段语音信号的时域波形图，其中，横轴表示时间，纵轴表示信号在采样点的振幅。情感语音的时域分析就是在时域波形图的基础上，计算语音信号的短时特征量。本文采用时域分析提取的情感语音特征包括短时平均能量和基音频率。

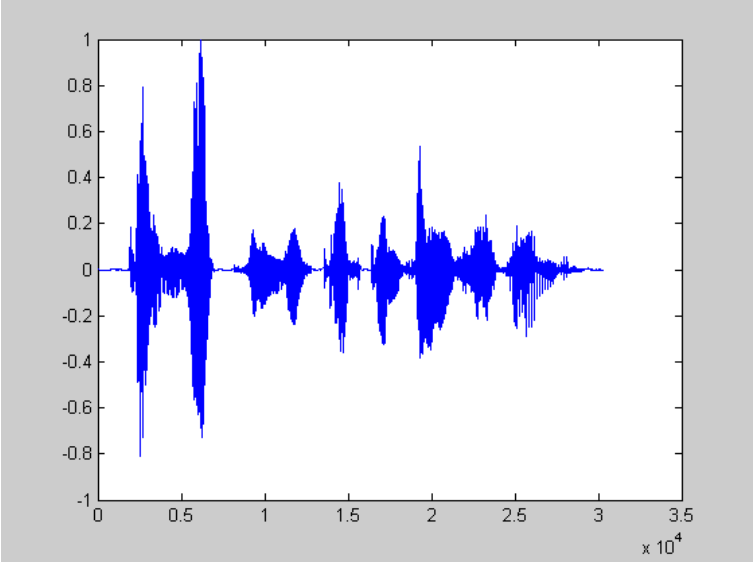

图 3-8 语音信号的时域波形图

浊音信号具有较强的振幅，且具有准周期性；清音信号振幅较弱，且不具有周期性。因此，浊音信号的短时平均能量较大，且有一定的基音频率；清音的短时平均能量较小，且没有基音频率。

3.5.1 短时平均能量

情感语音的能量特征反映了语音信号的振幅变化。由于语音信号具有短时平稳性，因此，可以在语音信号的每一个分析帧上分别计算短时能量特征，即短时平均能量。

短时平均能量定义为语音信号振幅的平方与窗函数平方的卷积，计算公式为

$$E_n = x^2(n) * \omega^2(n) = \sum_{m=-\infty}^{+\infty} x^2(m)\omega^2(n-m).$$

考虑窗口长度，短时平均能量的计算公式可写作

$$E_n = \sum_{m=n}^{n+(N-1)} x^2(m)\omega^2(n-m), \tag{3.5}$$

其中，N 是窗函数的长度。选取合适的窗口长度对短时平均能量特征的计算有较大影响。因为，若窗口长度很大，则窗函数相当于很窄的低通滤波器，E_n 随时间变化很小，能量特征变化的细节被忽略；相反，若窗口长度很小，则窗函数

相当于很宽的低通滤波器，E_n 随时间变化很大，不能得到比较平滑的能量特征。

图 3-9 为上文图 3-8 中语音信号在不同窗口长度下的短时平均能量。

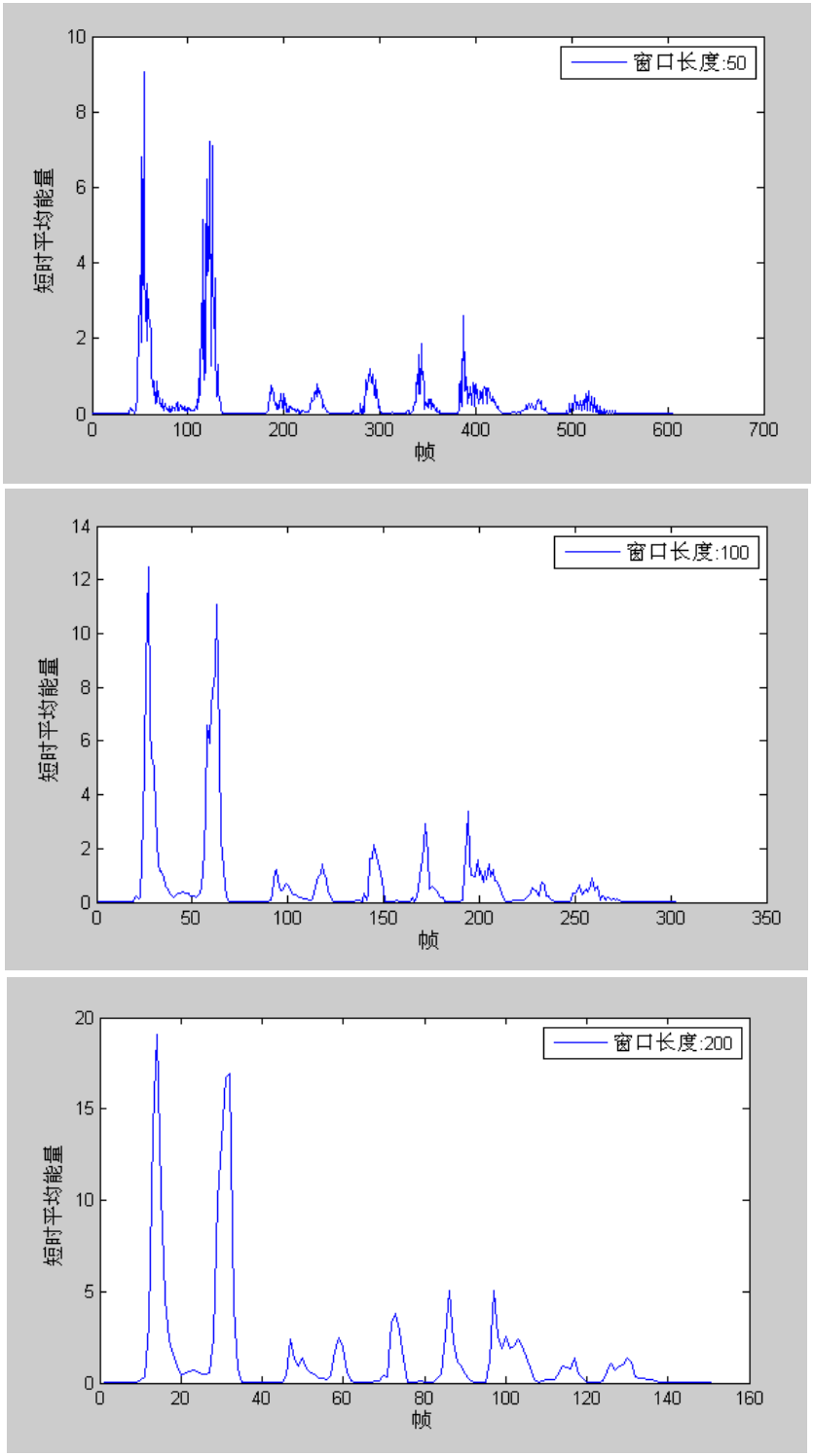

图 3-9 语音信号的短时平均能量

利用语音信号的短时平均能量特征，可以区分有声段和无声段，区分浊音和清音。这是因为语音信号有声段的短时平均能量远高于无声段，浊音相比清音具有较高的短时平均能量。

短时平均能量对于不同的语音情感也有一定的区分性。通常，同一个主体在表达高兴、惊奇、愤怒等情感时，语音的短时平均能量较大；而在表达悲伤情感时，短时平均能量较小。

类似于短时平均能量，短时平均幅值也能反映语音信号的振幅和能量变化。其定义为语音信号振幅的绝对值与窗函数的卷积，计算公式为

$$E_n = \sum_{m=n-(N-1)}^{n} |x(m)|\omega(n-m).$$

3.5.2 短时自相关函数

情感语音的自相关函数反映语音信号自身时域波形的相似性。根据语音信号的短时平稳性，在每一帧上分别计算信号的自相关函数，可得短时自相关函数。利用短时自相关函数，可以计算出语音信号的基音周期和基音频率。

相关函数用于衡量信号在时域的相似性，分为互相关函数和自相关函数。其中，互相关函数表达了两组信号之间的相关性，而自相关函数表达了一组信号自身的相似性和周期性。

信号的自相关函数计算公式为

$$r(k) = \sum_{m=-\infty}^{+\infty} x(m)x(m+k).$$

若信号具有周期性，则自相关函数具有与信号相同的周期，在信号周期的整数倍处，自相关函数达到极大值。自相关函数的这一性质为计算信号的周期提供了途径，即可以根据自相关函数极大值点之间的距离估算信号的周期。

考虑语音信号的短时平稳性和窗口长度，短时自相关函数的计算公式可写作

$$r_n(k) = \sum_{m=n}^{n+(N-1-k)} [x(m)\omega(n-m)][x(m+k)\omega(n-(m+k))].$$

窗口长度 N 应大于一个基音周期，否则将难以找出距离 $r(0)$ 最近的极大值点。将窗口长度由 N 增大到 $N+k$，可得到修正的短时自相关函数

$$\hat{r}_n(k) = \sum_{m=n}^{n+(N-1)} [x(m)\omega(n-m)][x(m+k)\omega(n-(m+k))]. \tag{3.6}$$

图 3-10 为一帧浊音信号短时自相关函数及其修正函数。由图可知，随着 k 的增大，短时自相关函数的幅度渐渐衰减，而修正的短时自相关函数幅度变化不明显。对于浊音信号，其短时自相关函数或修正函数具有与信号相同的周期，且函数在

信号周期的整数倍处达到极大值；清音的短时自相关函数不具有周期性，其幅度随 k 的增大迅速地减小。

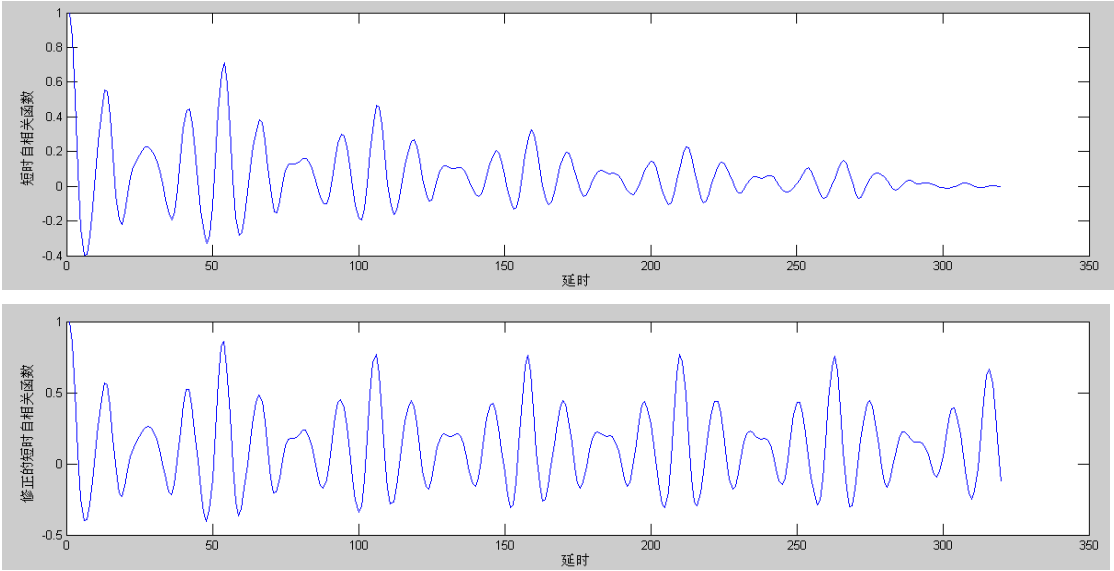

图 3-10 语音信号的短时自相关函数

类似于短时自相关函数，短时平均幅度差函数也能反映语音信号自身时域波形的相似性。短时平均幅度差函数的计算公式为

$$\gamma_n(k)=\sum_{m=n}^{n+(N-1-k)}|x(m+k)\omega(n-(m+k))-x(m)\omega(n-m)|.$$

对于浊音信号，其短时平均幅度差函数具有与信号相同的周期，在信号周期的整数倍处，函数急速降至波谷。因此，利用短时平均幅度差函数也可估算浊音信号的基音频率，且避免了计算复杂度较高的乘法运算。

3.5.3 基音频率

情感语音的基音频率反映了语音信号的准周期性。声带振动时，声带每开启和闭合一次的时间称为基音周期，基音周期的倒数为基音频率。由于语音信号具有短时平稳性，因此，可以在语音信号的每一个分析帧上分别计算基音频率。浊音信号具有准周期性，可以计算其基音频率；清音没有基音频率。

计算语音信号基音频率主要有三种方法，分别是自相关法、倒谱法和线性预测分析法。其中，自相关法估算基音频率属于时域分析方法。

自相关法计算一段浊音信号基音频率的主要步骤如下。

1. 采用中心削波法对语音信号作预处理。对信号作中心削波，可以突出基音周期整数倍处的语音信号的振幅。中心削波函数为

$$f(x)=\begin{cases} x-x_L & (x>x_L) \\ 0 & (-x_L\le x\le x_L) \\ x+x_L & (x<-x_L) \end{cases}.$$

对上文图 3-8 中语音信号的第 20001 至第 21000 个采样点作中心削波处理，结果如图 3-11 所示。

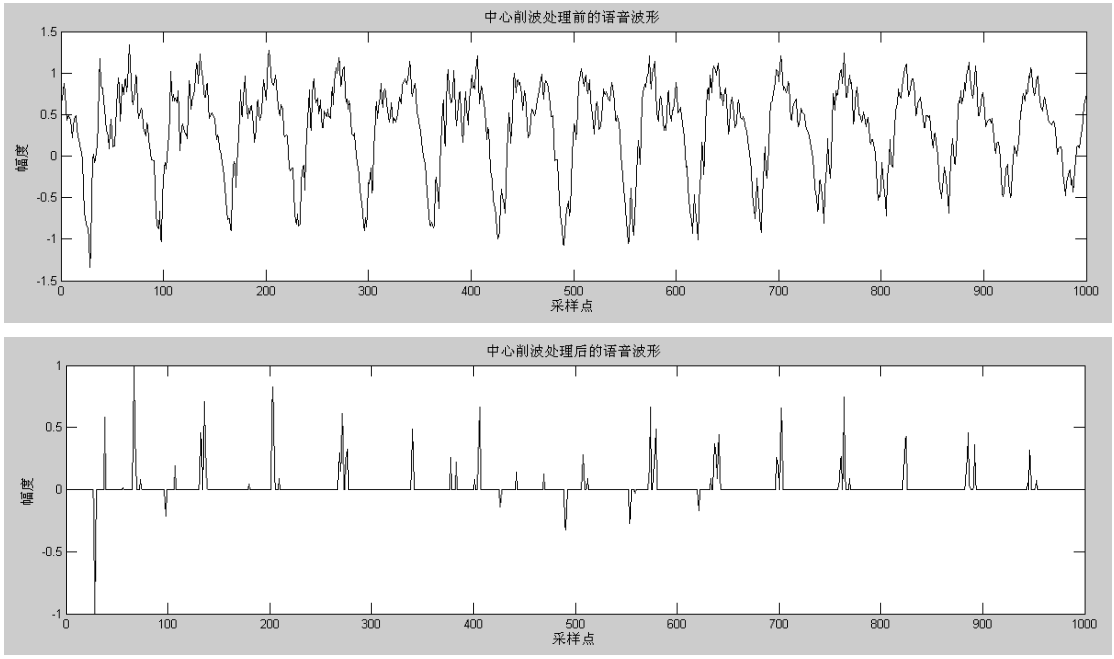

图 3-11 语音信号的中心削波处理

2. 计算每帧语音信号修正的短时自相关函数。图 3-12 为帧长为 320 时第一帧语音信号修正的短时自相关函数。

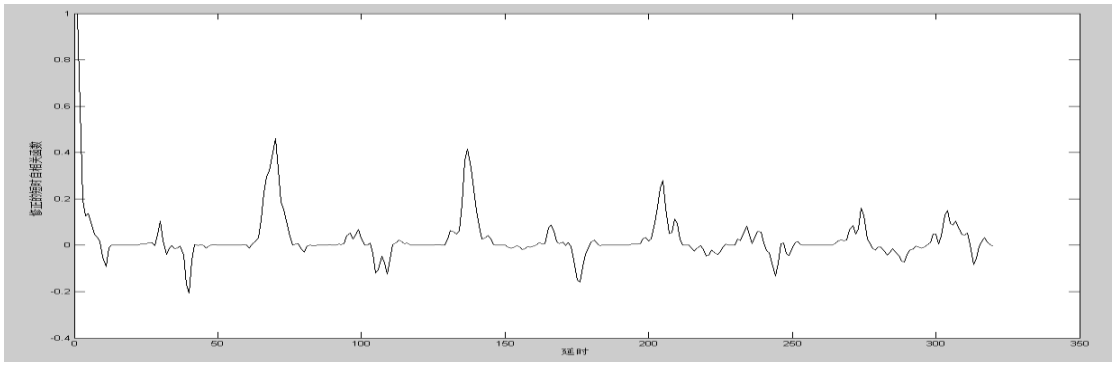

图 3-12 语音信号的短时自相关函数

3. 求出每帧信号自相关函数除原点外的第一个极大值点，其坐标即为这一帧语音信号的基音周期，基音周期的倒数是基音频率。

4. 对一段语音信号的基音频率轨迹作平滑处理，去除噪声。

利用基音频率能够较好地区分不同情感的语音信号。通常，高兴、愤怒、惊奇情感下的语音具有较高的基音频率，悲伤情感下的语音基音频率较低。

3.6 情感语音的频域分析

尽管语音信号的时域波形很直观，然而它比较容易受到环境因素的影响。相比之下，语音信号的频域波形受环境影响较小，且通过频域分析可以得到具有物理意义的语音特征，如基音频率、共振峰等。

傅里叶分析法是语音信号频域分析的常用方法。傅里叶分析法的基础是傅里叶变换，由法国人 J. Fourier 在十九世纪初提出。

周期信号可以展开成傅里叶级数。设信号 $x(t)$ 的周期为 T，则角频率 $\omega_0 = 2\pi / T$. 若 $x(t)$ 满足狄里赫利条件，则可以展开成傅里叶级数。傅里叶级数包括三种形式。第一种为三角形式

$$x(t) = a_0 + \sum_{n=1}^{+\infty}(a_n \cos n\omega_0 t + b_n \sin n\omega_0 t).$$

其中，$a_0 = \frac{1}{T}\int_{-\frac{T}{2}}^{\frac{T}{2}} x(t)dt$，$a_n = \frac{2}{T}\int_{-\frac{T}{2}}^{\frac{T}{2}} x(t)\cos n\omega_0 t dt$，$a_n = \frac{2}{T}\int_{-\frac{T}{2}}^{\frac{T}{2}} x(t)\sin n\omega_0 t dt$. 第二种也为三角形式

$$x(t) = a_0 + \sum_{n=1}^{+\infty} A_n \sin(n\omega_0 t + \varphi_n).$$

其中，$A_n = \sqrt{a_n^2 + b_n^2}$ 称为 $x(t)$ 的幅度频谱或频谱，$\varphi_n = \arctan\frac{a_n}{b_n}$ 称为 $x(t)$ 的相位频谱。第三种为指数形式

$$x(t) = \sum_{n=-\infty}^{+\infty} C_n e^{jn\omega_0 t}.$$

其中，$C_n = \frac{1}{T}\int_{-\frac{T}{2}}^{\frac{T}{2}} x(t)e^{-jn\omega_0 t}dt$ 称为 $x(t)$ 的幅度频谱或频谱。周期信号频谱的谱线是离散的，且谱线在基波频率的整数倍上。

随着周期 T 增大，频谱的谱线越来越密集。当 $T \to +\infty$ 时，谱线无限密集，频谱由离散变成连续。非周期信号可以认为是 $T \to +\infty$ 的周期信号, 非周期信号的谱线是连续的。采用极限的方法，可以从周期信号的频谱推导出非周期信号的频谱。非周期信号的频谱称为傅里叶变换，其表达式为

$$X(\omega) = \int_{-\infty}^{+\infty} x(t)e^{-j\omega t}dt.$$

傅里叶反变换的表达式为

$$x(t)=\frac{1}{2\pi}\int_{-\infty}^{+\infty}X(\omega)e^{j\omega t}d\omega .$$

由于语音信号具有短时平稳性，因此，可以对信号在每一帧上作傅里叶变换，称为短时傅里叶变换。考虑窗口长度，短时傅里叶变换的计算公式为

$$X_n(e^{j\omega})=\sum_{m=n}^{n+(N-1)}x(m)\omega(n-m)e^{-j\omega m} . \tag{3.7}$$

通过短时傅里叶变换得到一帧语音信号的频谱，利用频谱可以找出语音信号的基音频率和共振峰。图 3-13 是上文图 3-8 中帧长为 320 时，语音信号第 82 帧的时域波形图；图 3-14 是经傅里叶变换后得到的对数频谱图。由对数频谱曲线的包络可以看出，这一帧语音信号的基音频率大约是 100Hz, 前三个共振峰大约是 350Hz、1000Hz 和 1800Hz.

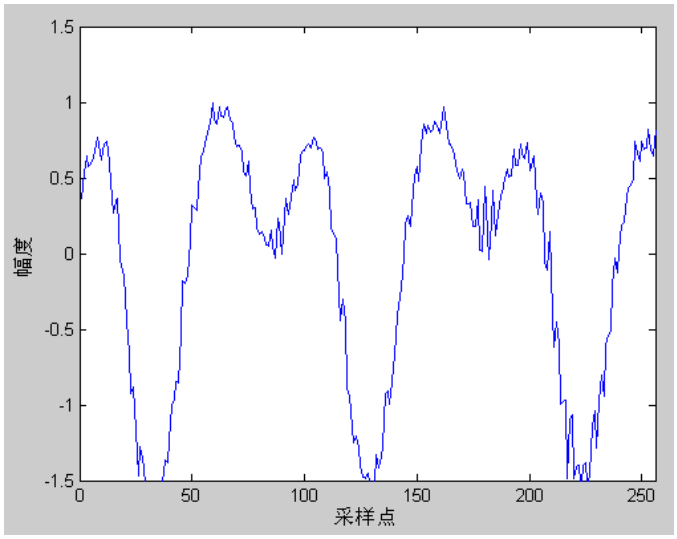

图 3-13 语音信号的时域波形图

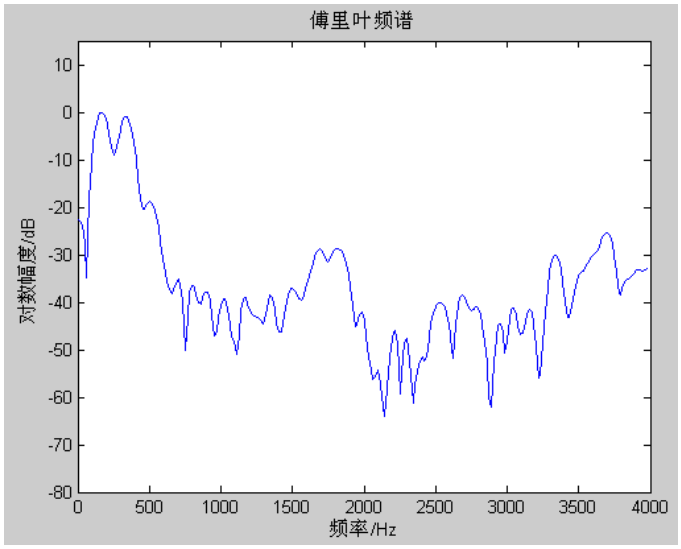

图 3-14 语音信号的傅里叶频谱

短时傅里叶变换频谱的平方是短时功率谱 $P_n(e^{j\omega})$. 短时功率谱也可通过短时自相关函数作傅里叶变换得到。一段语音信号的短时功率谱是一个二元函数。以时间为横轴、频率为纵轴构成二维平面，将短时功率谱按灰度描绘在对应坐标上，得到的图像称为语谱图。图 3-15 为上文图 3-8 中语音信号的语谱图。

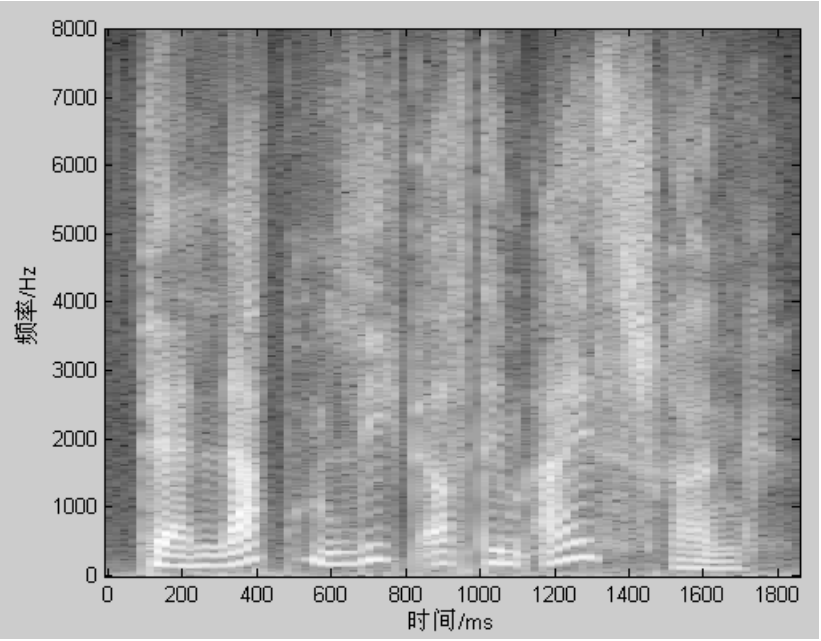

图 3-15 语谱图

功率谱的对数，称为对数功率谱。对对数功率谱作傅里叶反变换，可得一段语音信号的倒谱。在倒谱域分析语音信号，也可以计算出信号的基音频率和共振峰等特征量，且相比于时域分析和频域分析，特征量提取的准确度更高。

如图 3-16 所示为上述几组傅里叶变换谱之间的关系。由图可知，语音的时域波形经过傅里叶变换转化为傅里叶频谱，傅里叶频谱开平方得到功率谱，语音的自相关函数经过傅里叶变换转化成功率谱，功率谱取对数为对数功率谱，对数功率谱经过傅里叶反变换转化成倒谱；另外，在相反的方向上，傅里叶频谱经过傅里叶反变换转化为语音的时域波形，功率谱经过傅里叶反变换转化成语音的自相关函数，倒谱经过傅里叶变换转化为对数功率谱。

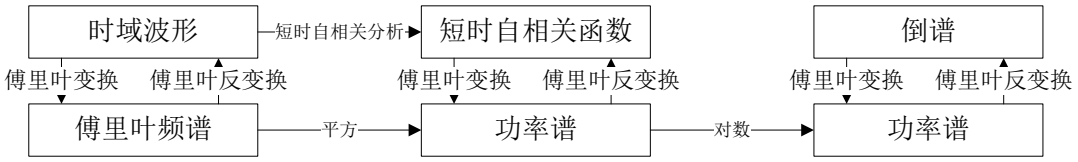

图 3-16 傅里叶变换谱

3.7 本章小结

本章分别从面部图像序列和情感语音信号中提取表情特征和语音特征。采用

面部动画参数作为表情特征，面部动画参数的计算基于面部特征点在图像序列中的位移，利用主动外观模型可以实现面部特征点的定位和跟踪。对语音信号作时域和频域分析，提取各帧的短时平均能量、基音频率和共振峰作为语音特征。

第四章 融合表情和语音的情感识别

基于隐马尔可夫模型和多层感知器，设计融合表情和语音特征的情感识别算法。利用提取的表情和语音特征，采用 Viterbi 算法训练各种表情和语音情感的隐马尔可夫模型；利用特征向量关于各隐马尔可夫模型的条件概率，采用反向传播学习算法训练多层感知器。

4.1 表情和语音 HMM 的拓扑结构

隐马尔可夫模型适合于处理与时间相关的问题。若样本是对应于连续时刻的特征向量，则可用来训练隐马尔可夫模型。本文从面部表情图像序列中提取的面部动画参数特征向量，以及从情感语音信号中提取的包含短时平均能量、基音频率和共振峰等语音情感特征的向量，都是对应连续时刻的特征向量。因此，可以利用表情和语音情感特征向量训练各种表情和语音情感的隐马尔可夫模型。

在基本的马尔可夫模型中，某时刻发生某事件的概率仅受到前一时刻发生的事件影响。基本的马尔可夫模型状态集为 $S=\{s_i\},i=1,2,\cdots,N$，其中， N 是状态数；初始时刻的概率分布为 $\pi=\{\pi_i\},i=1,2,\cdots,N$，其中， π_i 为初始时刻系统处于状态 i 的概率， $\sum_{i=1}^{N}\pi_i=1$；状态转移概率矩阵 $A=\{a_{ij}\},i=1,2,\cdots,N,j=1,2,\cdots,N$，其中， a_{ij} 为系统在某一时刻处于状态 i，而在下一时刻处于状态 j 的概率，且有 $\sum_{j=1}^{N}a_{ij}=1,i=1,2,\cdots,N$.

图 4-1 为基本的马尔可夫模型的图示。图中的结点表示状态，有向边表示状态转移概率。由于系统某一时刻的状态可以与下一时刻的状态相同，因此可能有 $a_{ij}\neq 0$；由于转移概率矩阵不一定是对称的，所以不一定有 $a_{ij}=a_{ji}$.

根据马尔可夫模型的定义，可以计算出系统产生某状态序列的概率。设马尔可夫模型为 λ，状态序列为 $Q=q_1q_2\cdots q_T$，则系统产生状态序列 Q 的概率

$$P(Q\mid\lambda)=\pi_{q_1}a_{q_1q_2}a_{q_2q_3}\cdots a_{q_{T-1}q_T}.$$

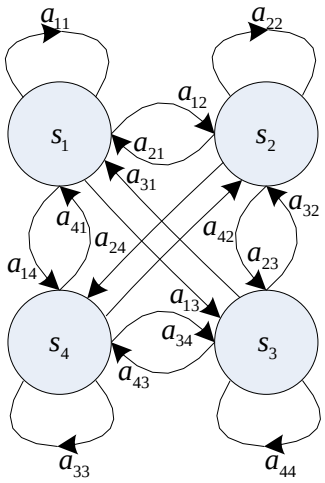

图 4-1 马尔可夫模型

基本的马尔可夫模型中的状态不是可观察的，即它们是抽象的概念，而不是具体的可观察的样本。比如，某种表情的马尔可夫模型有三个状态，分别表示表情处于平静、表情逐渐增强以及表情处于峰值；图像采集系统并不能直接观察到这些状态，而只能拍摄到不同时刻的面部表情图像，并由此计算出表征面部特征点位移的面部动画参数作为表情特征。因此，为了能够利用训练样本构建模型，需要在基本的马尔可夫模型中引入观察集，表示可观察的特征向量，由此可得隐马尔可夫模型。

隐马尔可夫模型在马尔可夫模型的基础上引入了观察集，包含全部的观察向量；状态集中的每个状态都有与观察向量相关的概率分布。隐马尔可夫模型的状态集为 $S=\{s_i\},i=1,2,\cdots,N$，其中，N 是状态数；观察集为 $V=\{v_i\},i=1,2,\cdots,M$，其中，M 是观察向量数；初始时刻的概率分布为 $\pi=\{\pi_i\},i=1,2,\cdots,N$，其中，$\pi_i$ 为初始时刻系统处于状态 i 的概率，$\sum_{i=1}^{N}\pi_i=1$；状态转移概率矩阵 $A=\{a_{ij}\},i=1,2,\cdots,N,j=1,2,\cdots,N$，其中，$a_{ij}$ 为系统在某一时刻处于状态 i，而在下一时刻处于状态 j 的概率，且有 $\sum_{j=1}^{N}a_{ij}=1,i=1,2,\cdots,N$；状态概率分布矩阵 $B=\{b_j(v_k)\},j=1,2,\cdots,N,k=1,2,\cdots,M$，其中，$b_j(v_k)$ 为系统处于状态 j 时观察向量为 v_k 的概率，且有 $\sum_{k=1}^{M}b_{ik}=1,i=1,2,\cdots,N$. 图 4-2 为隐马尔可夫模型的拓扑结构，图中的结点表示状态，连接结点的有向边表示状态转移概率，结点发出的有向边

表示状态观察概率。

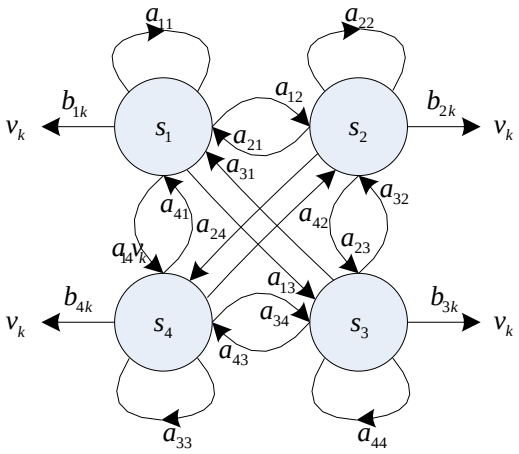

图 4-2 隐马尔可夫模型

上图中的隐马尔可夫模型状态间的转移是任意的；本文采用的隐马尔可夫模型是从左向右传递的隐马尔可夫模型，状态不是可遍历的，如图 4-3 所示，若某一时刻系统的状态为 S_i，则下一时刻的状态只能为 S_i 或 S_{i+1}，即状态只能保持不变，或转移到下一个相邻的状态。

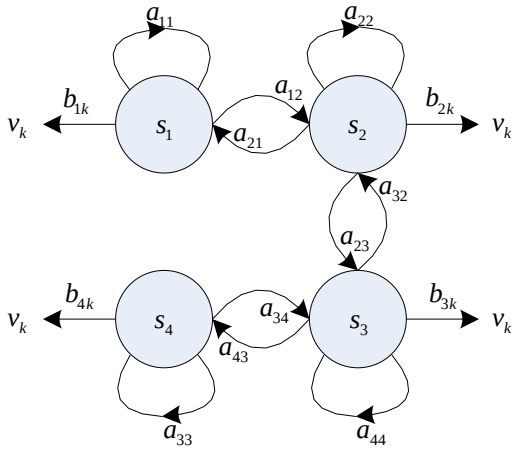

图 4-3 从左向右传递的隐马尔可夫模型

根据隐马尔可夫模型的定义，可以计算出系统产生某状态序列的概率。设隐马尔可夫模型为 λ，若状态序列为 $Q=q_1q_2\cdots q_T$，则系统产生状态序列 Q 的概率

$$P(Q|\lambda)=\pi_{q_1}a_{q_1q_2}a_{q_2q_3}\cdots a_{q_{T-1}q_T}; \tag{4.1}$$

若观察序列为 $O=o_1o_2\cdots o_T$，则系统状态序列为 Q 时，观察序列为 O 的概率

$$P(O|Q,\lambda)=b_{q_1}(o_1)b_{q_2}(o_2)\cdots b_{q_T}(o_T). \tag{4.2}$$

4.2 训练表情和语音的 HMM 模型

训练表情和语音情感 HMM 模型的样本分别为从面部表情序列中提取的表情特征以及从情感语音信号中提取的情感语音特征。具体地，表情特征为根据面部特征点位移计算出的面部动画参数，情感语音特征包括短时平均能量、基音频率和共振峰。表情特征和情感语音特征分别构成隐马尔可夫模型的观察向量，用于训练模型参数。本文训练的 HMM 模型为连续的 HMM 模型。

4.2.1 混合高斯分布的连续 HMM 模型

根据观察向量的分布，将隐马尔可夫模型分为离散的和连续的两种。若观察向量的分布是离散的，则训练出的 HMM 模型为离散 HMM 模型；若观察向量的分布是连续的，则训练出的 HMM 模型为连续 HMM 模型。通常连续 HMM 模型观察向量符合高斯分布或混合高斯分布，即观察向量的概率密度函数为多元正态密度函数或多元正态密度函数的加权求和。

高斯分布的概率密度函数为

$$p(x)=\frac{1}{(2\pi)^{d/2}|\sum|^{1/2}}\exp[-\frac{1}{2}(x-\mu)^T\sum^{-1}(x-\mu)],$$

其中，μ 为高斯分布的均值向量，$\sum$ 为协方差矩阵，记作 $p(x)\sim N(\mu,\sum)$.

混合高斯分布的概率密度函数为

$$p(x)=\sum_{i=1}^{m}c_i N(\mu_i,\sum_i),$$

其中，m 为高斯混合数，c_i 为第 i 个高斯混合的权重。

高斯分布和混合高斯分布中的参数可以采用最大似然估计法计算出。若 x_i 为符合高斯分布的样本，其中，$i=1,2,\cdots,n$, 则高斯分布的均值向量和协方差矩阵的最大似然估计结果为

$$\hat{\mu}=\frac{1}{n}\sum_{i=1}^{n}x_i,$$

$$\hat{\sum}=\frac{1}{n}\sum_{i=1}^{n}(x_i-\hat{\mu})(x_i-\hat{\mu})^T.$$

本文用于训练 HMM 模型的观察向量分别为表情特征和情感语音特征，可以认为它们符合混合高斯分布。训练 HMM 模型的主要步骤包括模型的初始化、状

态的 Viterbi 分割以及模型参数的 Baum-Welch 重估计。

4.2.2 基于 K-Means 聚类的模型初始化

根据以上的分析，本文采用的 HMM 模型为从左向右传递的连续 HMM 模型，观察向量符合混合高斯分布。为训练 HMM 模型，首先需要初始化模型参数。

对训练集中全部观察向量序列的状态序列作均一分割。设 HMM 模型的状态数为 N，若某观察向量序列的长度为 T，则该序列前 $\lfloor T/N \rfloor$ 个观察向量处于状态 1，第 $\lfloor T/N \rfloor+1$ 个到第 $2\lfloor T/N \rfloor$ 个观察向量处于状态 2，…，第 $(N-1)\lfloor T/N \rfloor+1$ 个到第 T 个观察向量处于状态 N。

采用 K-Means 聚类算法，将各状态的观察向量分配给不同的高斯混合。假设某状态有 m 个高斯混合，即该状态下观察向量的概率密度函数为 m 个多元正态密度函数的加权和。对训练集中该状态下全部观察向量按 K-Means 算法作聚类，得到 m 个聚类，属于各聚类的观察向量即为分配给该状态各高斯混合的观察向量，聚类中心即为各高斯混合的均值向量。

设属于某高斯混合的观察向量数为 N_i，该高斯混合所属状态下的全部观察向量数为 N，则该高斯混合的权重为 $c_i = \frac{N_i}{N}$；利用属于高斯混合的观察向量可以计算出该高斯混合的均值向量和协方差矩阵。由于本文采用连续 HMM 模型的观察向量符合混合高斯分布，根据某状态下各高斯混合的权重、均值向量和协方差矩阵，可得该状态下观察向量的概率密度函数

$$b_j(o_k) = \sum_{i=1}^{m} \frac{c_i}{(2\pi)^{d/2} |\sum_i|^{1/2}} \exp[-\frac{1}{2}(o_k - \mu_i)^T \sum_i^{-1} (o_k - \mu_i)], \tag{4.3}$$

其中，$b_j(o_k)$ 为系统处于状态 j 时观察向量为 o_k 的概率，m 为状态 j 包含的高斯混合数，c_i 为第 i 个高斯混合的权重，μ_i 为均值向量，$\sum_i$ 为协方差矩阵。

4.2.3 基于 Viterbi 算法的状态分割

初始化 HMM 模型参数后，利用 Viterbi 算法对训练集中观察向量序列的状态序列重新作分割，直至模型参数不再发生明显变化。

已知 HMM 模型 λ，观察向量序列 O，利用 Viterbi 算法可以计算出使 $P(O,Q|\lambda)$ 最大的状态序列 Q.

由于

$$P(O,Q|\lambda)=P(O|Q,\lambda)P(Q|\lambda)=\prod_{t=1}^{T}b_{q_t}(o_t)\pi_{q_1}\prod_{t=1}^{T}a_{q_{t-1}q_t}=\pi_{q_1}\prod_{t=1}^{T}a_{q_{t-1}q_t}b_{q_t}(o_t),$$

定义局部概率

$$\varphi_j(t)=\begin{cases}\pi_j b_j(o_t), t=1\\ \max\limits_{i=1}^{N}\{\varphi_i(t-1)a_{ij}b_j(o_t)\}, t=2,3,\cdots,T\end{cases},$$

其中 $j=1,2,\cdots,N$，则 $\max\limits_{i=1}^{N}\varphi_i(T)$ 为 $P(O,Q|\lambda)$ 的最大值。设使 $P(O,Q|\lambda)$ 最大的状态序列为 $Q=q_1q_2\cdots q_T$，则 $q_i=\arg\max\limits_{j}\{\varphi_j(t)\}$.

利用 Viterbi 算法训练 HMM 模型参数的迭代算法如下。

1. 对训练集中观察向量序列的状态序列作均一分割。

2. 采用 K-Means 聚类算法，将各状态的观察向量分配给不同的高斯混合。

3. 根据状态包含的各高斯混合的权重、均值向量和协方差矩阵，计算各状态下观察向量的概率密度函数。

4. 对训练集中观察向量序列的状态序列作 Viterbi 分割。

5. 根据观察向量到各高斯混合均值的欧氏距离，重新将各状态的观察向量分配给不同的高斯混合。

6. 重复 3-5，直至相邻两次迭代得到的全部 $P(O,Q|\lambda)$ 的最大值之和小于一定的阈值。

4.2.4 Baum-Welch 参数重估计

利用 Baum-Welch 算法可对 Viterbi 算法训练得到的 HMM 参数作重估计。Baum-Welch 算法（或称为前向-后向算法）是广义期望最大化算法（即 EM 算法）的一种实现。Baum-Welch 算法利用前向算法和后向算法中的局部概率，迭代地更新HMM模型的状态转移概率 a_{ij} 和状态观察概率 $b_j(v_k)$ 等参数，直至达到收敛。

然而，仅利用 Viterbi 算法也可训练得到的近似的 HMM 模型参数，因此，不再采用 Baum-Welch 算法对参数作重估计。

4.3 基于 HMM 模型的表情和语音情感识别

利用训练出的表情或语音情感的隐马尔可夫模型，可以识别出面部表情序列图像测试样本或情感语音测试样本的情感状态。

提取测试样本的表情特征或情感语音特征，构成 HMM 模型的测试观察向量序列。设 HMM 模型为λ, 测试观察向量序列为O, 采用前向算法可以计算出使$P(O|\lambda)$的值.

根据

$$P(O|\lambda)=\sum_{k=1}^{k_{\max}}P(O|Q_k,\lambda)P(Q_k|\lambda)=\sum_{k=1}^{k_{\max}}\prod_{t=1}^{T}b_{q_{k,t}}(o_t)\pi_{q_{k,1}}\prod_{t=1}^{T}a_{q_{k,t-1}q_{k,t}}=\sum_{k=1}^{k_{\max}}\pi_{q_{k,1}}\prod_{t=1}^{T}a_{q_{k,t-1}q_{k,t}}b_{q_{k,t}}(o_t),$$

其中，Q_k为一种状态序列，$k_{\max}$为所有可能的状态序列数，定义局部概率

$$\alpha_j(t)=\begin{cases}\pi_j b_j(o_t), t=1\\ \sum_{i=1}^{N}\alpha_i(t-1)a_{ij}b_j(o_t), t=2,3,\cdots,T\end{cases},$$

其中，$j=1,2,\cdots,N$, 则$\sum_{i=1}^{N}\alpha_i(T)$为$P(O|\lambda)$的值.

运用局部概率计算$P(O|\lambda)$, 将大大降低计算复杂性。这是因为，若采用穷举法直接根据公式$P(O|\lambda)=\sum_{k=1}^{k_{\max}}P(O,Q_k|\lambda)$对所有可能的$Q_k$求和，则时间复杂度高达$O(N^TT)$；若采用局部概率公式迭代地计算$P(O|\lambda)$, 则时间复杂度陡降至$O(N^2T)$, 效率极大地提升。

根据贝叶斯公式$P(\omega_j|x)=\frac{p(x|\omega_j)P(\omega_j)}{p(x)}$, 其中，$x$是样本，$\omega_j$表示第$j$类，$P(\omega_j)$为先验概率，$p(x|\omega_j)$为类条件概率密度，$P(\omega_j|x)$为后验概率，可以得到$P(\lambda|O)=\frac{P(O|\lambda)P(\lambda)}{P(O)}$. 由此可得表情或语音情感的判别函数

$$g_i(O)=P(O|\lambda_i)P(\lambda_i), i=1,2,\cdots,L,$$

其中，λ_i为第i种表情或语音情感对应的 HMM 模型，L为表情或语音情感的种类数，$P(\lambda_i)$为第i种表情或语音情感的先验概率，先验概率根据各种情感状态下面部表情图像序列或情感语音的样本数量的先验知识得到，通常为均匀分布，可以忽略。计算测试样本对于不同情感的判别函数$g_i(O)$, 使$g_i(O)$达到最大值的情感状态即为测试样本情感状态。

为融合来自视频和音频的信息，需要保留观察向量序列关于各种情感状态HMM模型的条件概率$P(O|\lambda)$,用作融合策略的输入。本文采用人工神经网络在决策级融合多模态以识别情感，$P(O|\lambda)$可作为神经网络的输入值。

4.4 融合表情和语音 HMM 的多层感知器

在完成了表情特征和情感语音特征的提取，并分别为各种表情和语音情感建立了各自的HMM模型之后，可以对来自面部表情图像和语音两种模态的信息作融合。本文在决策级上融合模态，采用基于人工神经网络的融合策略，其中人工神经网络的训练样本是表情特征或情感语音特征组成的观察序列关于各 HMM模型的条件概率$P(O|\lambda)$,利用这些条件概率值可以训练出神经网络的权重矩阵，达到融合模态的目的。

人工神经网络模拟大脑的神经组织，以神经元为基本单元，具有网络拓扑结构和网络训练算法等要素，在一定程度上反映了神经组织的结构特点和认知过程。尽管每个神经元的结构简单、功能单一，但是大量的神经元组成的网络具有复杂的结构和强大的功能，使得人工神经网络具备较强的学习能力和容错性能。与大脑的功能相像，人工神经网络反映了各信息源之间的相互影响和相互制约，使来自不同模态的信息融合成一个有机的整体，因此是一种可以采用的多模态融合方法。根据拓扑结构和学习算法的不同，人工神经网络的模型分为很多种，分别可以应用于不同的场合。本文采用的人工神经网络模型为多层感知器。

多层感知器的基本单元是神经元。每个神经元都有若干个输入连接，它们使多层感知器上一层神经元的输出值成为该神经元的输入；也有若干个输出连接，它们向下一层神经元传递响应。如图 4-4 所示，多层感知器的神经元将从上一层取到的值加权求和，并加上偏置项，求和结果由激活函数作进一步处理。设$x_1, x_2, \cdots, x_N$是神经元的输入，$\omega_{j,1}, \omega_{j,2}, \cdots, \omega_{j,N}$是对应各输入的权值，$bias_j$是偏置项，$f(\cdot)$为激活函数，则神经元 j 的输出为

$$y_j = f(\sum_{i=1}^{N} x_i \omega_{j,i} + bias_j).$$

常用的神经元激活函数包括阶跃函数和 Sigmoid 函数。阶跃函数的表达式为

$$f(x) = \begin{cases} 1, x \geq 0 \\ 0, x < 0 \end{cases},$$

标准 Sigmoid 函数的表达式为

$$f(x)=\frac{1-e^{-x}}{1+e^{-x}}.$$

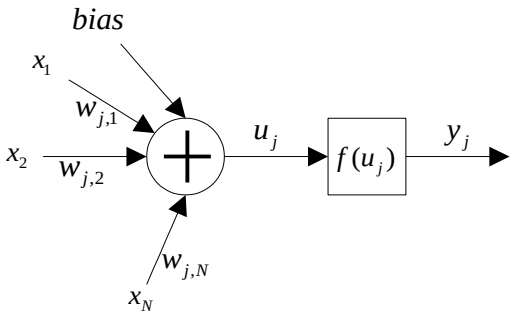

图 4-4 神经元结构

多层感知器由输入层、输出层和一个或多个隐藏层组成，每层包含若干神经元，各神经元与前后两层的神经元直接相连。图 4-5 为一个多层感知器的例子。图中的多层感知器包含一个输入层、一个隐藏层和一个输出层，其中，输入层有两个神经元，隐藏层有五个神经元，输出层有两个神经元。

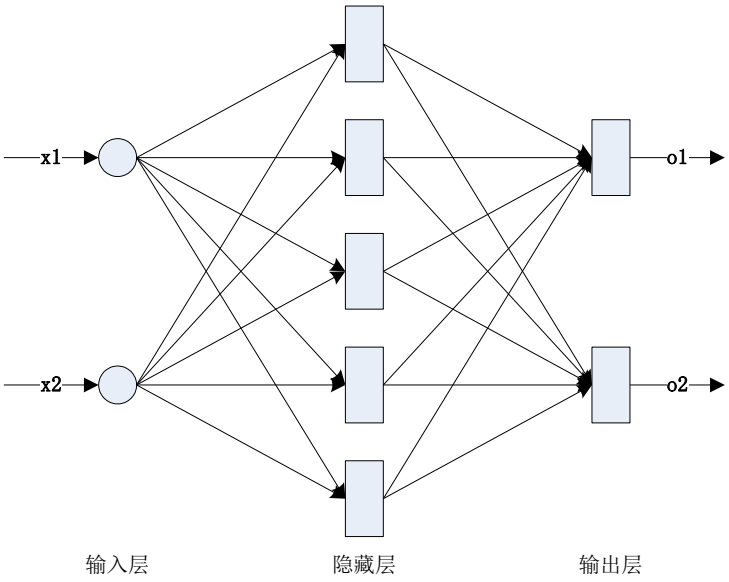

图 4-5 多层感知器的拓扑结构

4.5 训练融合表情和语音的多层感知器

为训练融合表情和情感语音的多层感知器，首先需要从面部表情图像序列和语音信号中提取情感特征，情感特征作为观察向量序列用于训练不同模态下不同情感的 HMM 模型，计算观察向量序列关于各 HMM 模型的条件概率 $P(O|\lambda)$，这些条件概率即为训练的多层感知器的输入；输出层的神经元数为模态融合后的

情感状态数。例如，某多层感知器用于融合四种表情状态和四种语音情感状态，则输入层有八个神经元，每个神经元的输出值分别为条件概率 $P(O|\lambda)$，输出层有四个神经元，每个神经元代表一种情感状态，由此可以训练出融合四种表情和语音情感状态的多层感知器。

多层感知器的训练采用反向传播学习算法。反向传播算法的训练过程分为两个阶段，分别是传播和更新权重。在对多层感知器实施反向传播算法之前，需要初始化权重矩阵和偏置项为小随机数。

传播阶段计算多层感知器的输出值。设多层感知器的输入为 $x_1, x_2, \cdots, x_N$，实际输出为 $z_1, z_2, \cdots, z_M$，其中，

$$z_i = \begin{cases} 1, \text{若训练样本属于第} i \text{类} \\ 0, \text{其它} \end{cases}, i = 1, 2, \cdots, M,$$

根据多层感知器的输入值、权重矩阵、偏置项以及激活函数，利用神经元输出的计算公式，逐层向后计算各神经元的输出值，直至得到多层感知器的输出 $y_1, y_2, \cdots, y_M$.

更新权重阶段从输出层逐层向前修正权重矩阵和偏置项。权重的修正公式为

$$\omega'_{j,i} = \omega_{j,i} + \eta \sigma_j x_i,$$

偏置项的修正公式为

$$bias'_j = bias_j + \eta \sigma_j .$$

其中，$\omega_{j,i}$ 为神经元 j 对应输入 x_i 的权重，$bias_j$ 为神经元 j 的偏置项，η 称为收敛因子，取 $0<\eta<1$，σ_j 称为校正因子。σ_j 的计算公式为

$$\sigma_j = \begin{cases} y_j(1-y_j)(z_j - y_j), \text{神经元} j \text{位于输出层} \\ y_j(1-y_j)\sum_{k=1}^{L} \sigma_k \omega_{k,j}, \text{神经元} j \text{位于隐藏层} \end{cases},$$

其中，y_j 为隐藏层或输出层神经元的输出值，σ_k 表示上一层（靠近输出层的相邻层）神经元的校正因子，L 为上一层的神经元数。

重复执行传播阶段和更新权重阶段的算法，直至权重矩阵和偏置项不再发生明显变化。

4.6 融合表情和语音的情感识别

测试样本为面部表情图像序列和相应的情感语音信号。从测试样本中提取表情特征和情感语音特征，计算情感特征关于各 HMM 模型的条件概率，利用训练得到的多层感知器的权重矩阵和偏置项，可以算得多层感知器的输出。由于多层感知器输出层的每个神经元代表一种情感状态，输出层最大输出值的神经元对应的情感状态即为模态融合后得到的测试样本的情感状态。

计算多层感知器的输出，可采用反向传播算法中传播阶段的算法，其中，神经元 j 对应上一层各神经元的权重 $\omega_{j,i}$，以及偏置项 $bias_j$ 都已经在多层感知器的训练过程中得到。

4.7 本章小结

本章设计算法实现了融合表情和语音特征的多模态情感识别。以表情和语音特征为观察向量序列，训练各种表情和语音的隐马尔可夫模型，计算观察向量序列关于各 HMM 模型的条件概率，作为多层感知器的输入；采用反向传播学习算法训练多层感知器的权重矩阵。对于测试样本，提取表情和语音特征，计算特征向量序列关于各 HMM 模型的条件概率，作为多层感知器的输入；计算多层感知器的输出，识别出测试样本的情感状态。

第五章 多模态融合情感识别实验

设计实验验证多模态融合情感识别算法。利用 Cohn-Kanade2010 人脸表情库、Berlin 情感语音库以及采集的视频，基于 HMM 模型和多层感知器，融合表情和语音特征识别情感。

5.1 实验目的

设计多模态情感识别实验，从面部图像序列中提取表情特征，从情感语音信号中提取语音的情感特征，基于隐马尔可夫模型和人工神经网络中的多层感知器，融合表情特征和语音的情感特征，识别人的情感状态。

5.2 实验数据库

实验主要运用了三种数据库，分别是采用 Cohn-Kanade2010 人脸库，Berlin 情感语音库，以及利用采集的情感视频短片建立的数据库。

5.2.1 面部表情数据库

实验采用 Cohn-Kanade2010 人脸库作为面部表情数据库，用于提取表情特征和训练表情 HMM 模型。

Cohn-Kanade2010 人脸表情数据库由 123 名志愿者的 593 组面部表情图像序列组成，每组表情序列从平静状态下的人脸图像开始，到表情达到峰值结束；每幅图像都是不变光照下、带有不同面部表情的正面人脸，图像格式为 PNG, 大小为 640*490（灰度图像）或 640*480（彩色图像）。图 5-1 为 CK2010 库中的面部表情图像。

图 5-1 Cohn-Kanade2010 人脸表情库

CK2010 库的每幅人脸图像带有一个标注文件，用于记录 68 个面部特征点在图像中的坐标位置。这些特征点分别来自眉、眼、鼻、外唇、内唇和下巴[39]。

另外，IMM 人脸数据库也可以作为训练集，为面部表情图像建立主动外观模型。

IMM 人脸数据库由 40 个人的 240 幅人脸图像组成，其中，7 人为女性，33 人为男性，图像是 640*480 像素的 JPEG 格式。

图 5-2 IMM 人脸库

每名志愿者拍摄的六幅图像分别代表了不同的光照强度、旋转角度和面部表情，它们的形式如下所述。

第一幅图像为正面人脸，平静表情，散射光线；

第二幅图像为正面人脸，高兴表情，散射光线；

第三幅图像中的人脸向右方旋转约 30 度角，中立表情，散射光线；

第四幅图像中的人脸向左方旋转约 30 度角，中立表情，散射光线；

第五幅图像为正面人脸，中立表情，人脸左侧有直射光照；

第六幅图像为正面人脸，志愿者可以做出任意的表情或姿态，散射光线。

图 5-2 中的六幅图像分别代表以上的六种图像。

IMM 库的每幅人脸图像配有一个.asf 格式的标注文件，标注文件记录了图像中 58 个特征点的在图像中的相对坐标位置。这些特征点分别取自眉、眼、鼻、嘴和下巴等面部器官，它们组成了七条路径，其中，三条路径闭合，分别构成眼和嘴的形状，四条路径不闭合，分别构成眉、鼻和下巴的形状[40]。

5.2.2 情感语音数据库

实验采用 Berlin 情感语音库作为语音数据库，用于提取语音特征和训练语音情感 HMM 模型。Berlin 情感语音库是德语语音数据库，十名志愿者用包含平静

情感的七种情感朗诵十句德文，每条语音时长一至三秒。十名志愿者中包含五名男性和五名女性[41]。这十句德文译成英文分别如下。

1. The tablecloth is lying on the fridge.
2. She will hand it in on Wednesday.
3. Tonight I could tell him.
4. The black sheet of paper is located up there besides the piece of timber.
5. In seven hours it will be.
6. What about the bags standing there under the table?
7. They just carried it upstairs and now they are going down again.
8. Currently at the weekends I always went home and saw Agnes.
9. I will just discard this and then go for a drink with Karl.
10. It will be in the place where we always store it.

5.2.3 情感视频采集

实验采集了带有表情和语音的视频短片，用于测试实验，验证算法的有效性。每段视频时长 3-5 秒，各帧图像的分辨率为 640*480 像素，视频中的志愿者做出高兴、悲伤、愤怒、惊奇等各种表情，并伴以带有对应情感的语音。图 5-3 为从采集的一段视频中截取的若干帧表情图像，这段视频表达了高兴情感。图 5-4 为从若干段视频中截取的表情峰值图像。

志愿者在完成表情的同时，以相同的情感状态朗诵一段文字，下面是其中五段文字的内容。

1. 今天天气可真好啊。
2. 这本书这么有意思。
3. 你要去那个地方啊。
4. 你居然认识他。
5. 这座校园可真大啊。

由于这些文字本身不带有任何感情色彩，因此，志愿者可以以高兴、悲伤、愤怒、惊奇等任意一种情感状态表达它们。

图 5-3 情感视频短片的图像序列

图 5-4 情感视频短片的表情峰值图像

5.3 实验设计

5.3.1 情感特征提取实验

实验提取的情感特征包括表情特征和语音的情感特征。表情特征即为根据检测的面部特征点坐标计算出的面部动画参数；语音的情感特征包括短时平均能量、基音频率和共振峰频率等。用于提取情感特征的数据库包括 Cohn-Kanade2010 人脸表情库，Berlin 情感语音库，以及自建的情感视频库。

为定位和跟踪面部特征点，需要建立面部表情图像的主动外观模型，采用迭代算法重建面部表情图像，并检测面部特征点坐标。用于训练主动外观模型的人脸图像取自 Cohn-Kanade2010 人脸表情库，训练图像中既有表情平静的图像，也有表情强烈的图像，图 5-5 为若干幅用于训练主动外观模型的人脸图像。

图 5-5 主动外观模型的训练图像

主动外观模型的训练图像都具有较强的代表性。训练集中的志愿者性别、种族各有不同，拍摄现场的光照强度也有所差异，志愿者的表情包括平静、高兴、愤怒、悲伤、厌恶、惊奇等等。采用这些图像训练出的主动外观模型，能够较好地反映人脸图像的差异；用于重建面部表情图像和检测面部特征点，具有更高的准确度。图 5-6 为面部表情图像重建，以及面部特征点检测的结果。图 5-7 为图

像序列中的面部特征点检测，图像序列取自采集的情感视频短片。

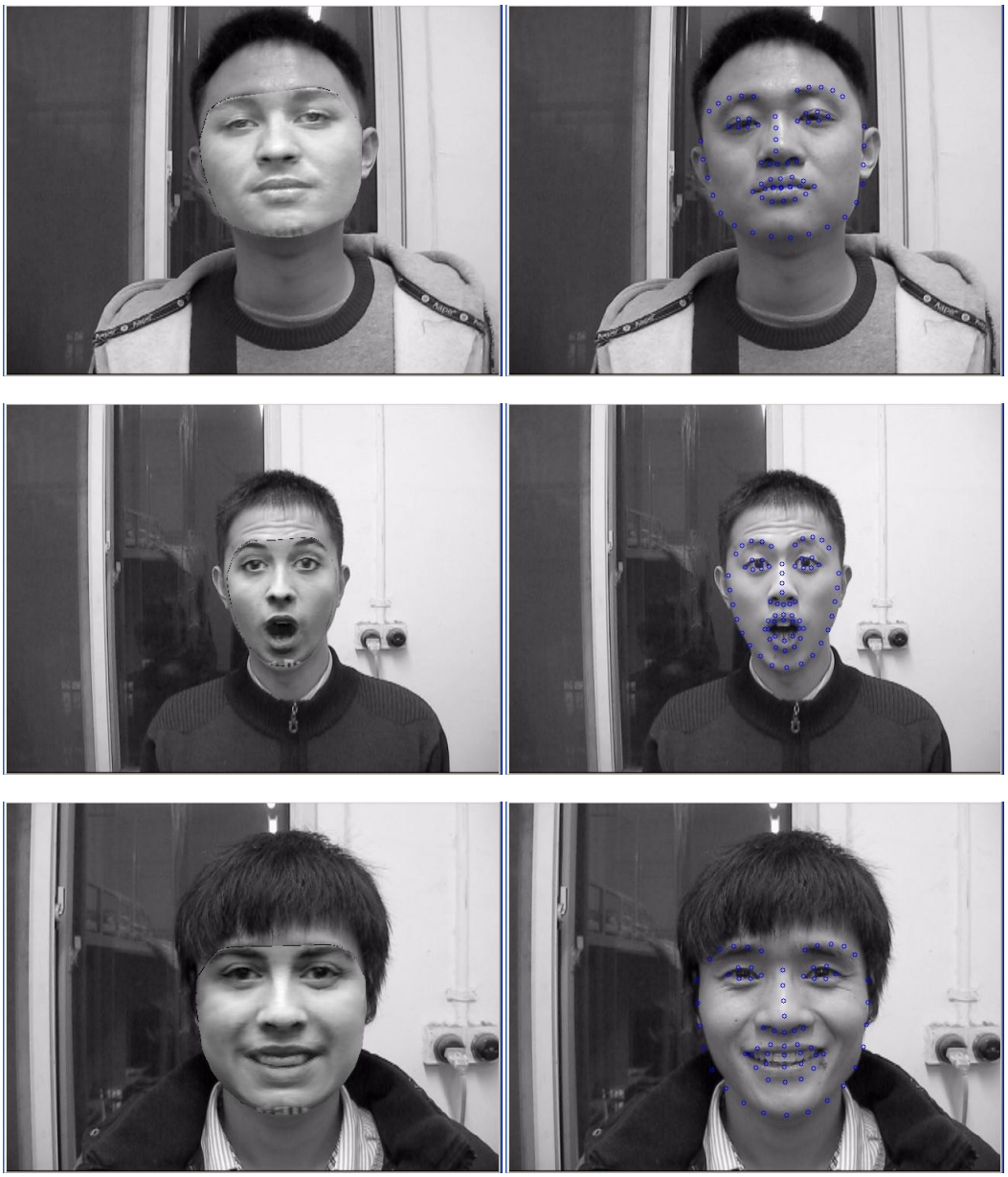

图 5-6 表情图像重建和面部特征点检测

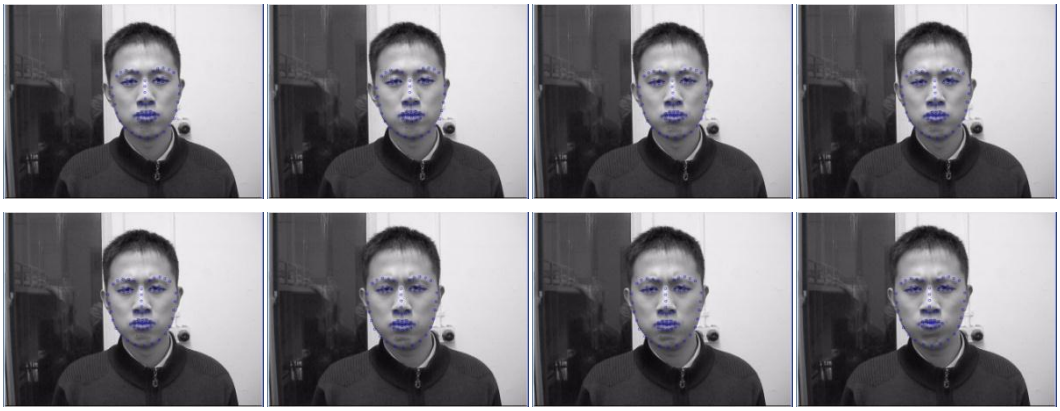

图 5-7 图像序列中的面部特征点定位和跟踪

根据图像序列中面部特征点的位移，计算出各帧图像的面部动画参数，作为表情特征。图 5-8 为高兴、悲伤、愤怒、厌恶四种表情对应的面部动画参数的示例，其中，横轴表示表情图像在图像序列中的帧号，纵轴表示面部动画参数值，各条曲线分别为 FAP 曲线，包括 FAP 3-13，19-22，31-38，51-60，61-64.

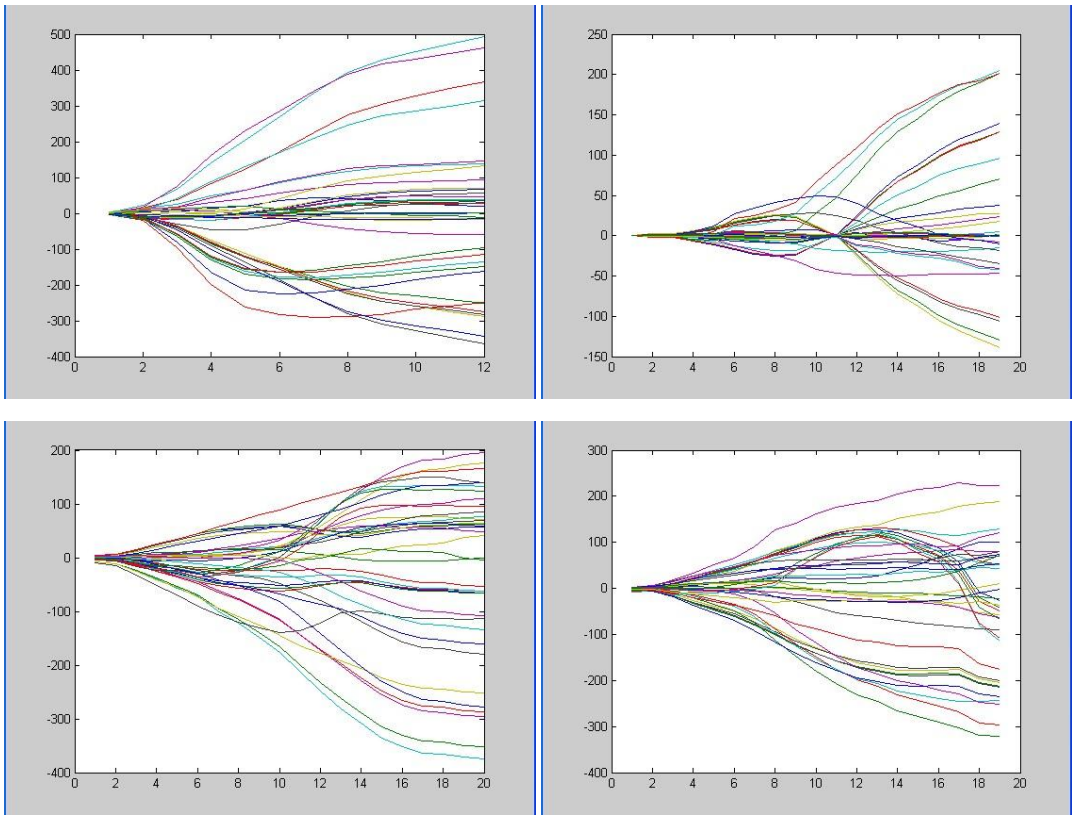

图 5-8 图像序列的面部动画参数

（左上：高兴表情；右上：悲伤表情；左下：愤怒表情；右下：厌恶表情）

对语音信号作时域和频域分析，可以计算出短时平均能量、基音频率和共振峰等语音的情感特征。图 5-9 为一段语音信号各帧的短时平均能量。图 5-10 为语音信号各帧的基音频率。图 5-11 为语音信号各帧的前四个共振峰。

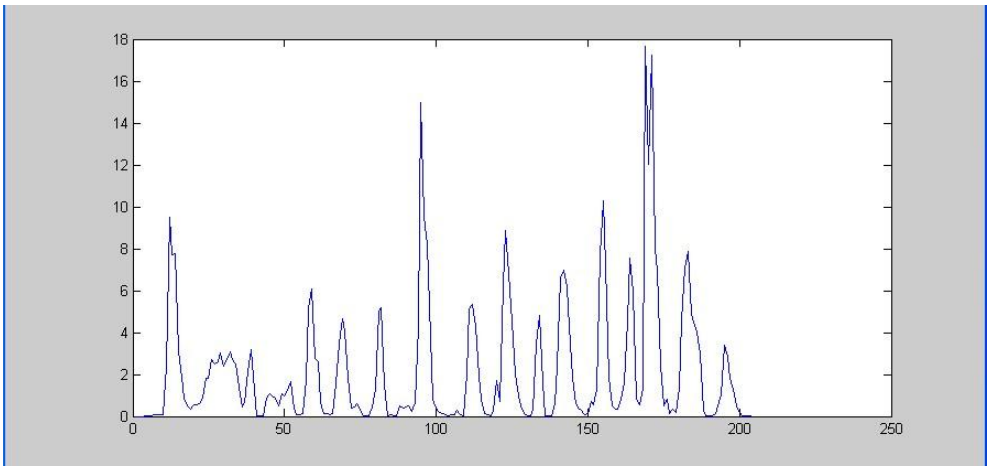

图 5-9 语音信号的短时平均能量

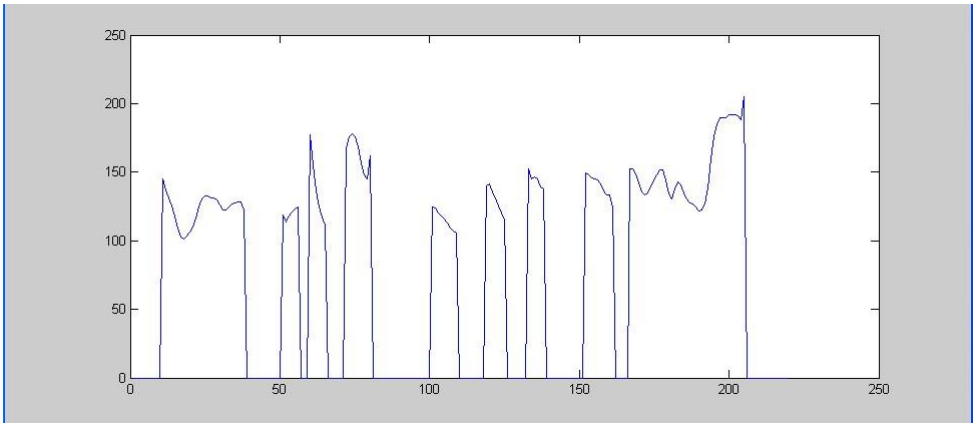

图 5-10 语音信号的基音频率

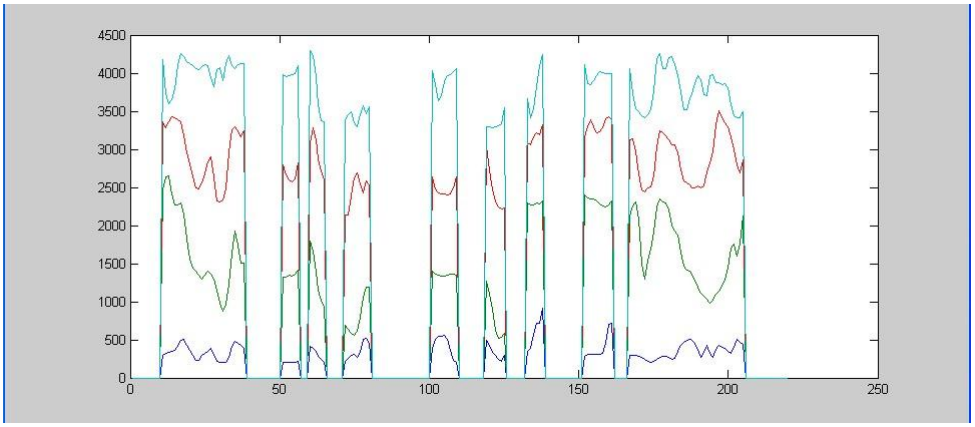

图 5-11 语音信号的前四个共振峰

5.3.2 多模态融合情感识别实验

利用从训练样本中提取的表情和语音特征构造 HMM 模型的观察向量序列，采用 Viterbi 算法训练各种表情和语音情感的 HMM 模型，具体流程如图 5-12 所示。

计算观察向量序列关于各 HMM 模型的条件概率，作为多层感知器的输入；采用反向传播学习算法训练多层感知器的权重矩阵。

从测试样本中提取表情和情感语音特征构成观察向量序列，计算观察向量序列关于各 HMM 模型的条件概率，作为多层感知器的输入；计算多层感知器的输出，识别出测试样本的情感状态。

用于训练和测试的表情图像序列样本来自 Cohn-Kanade2010 面部表情库，情感语音样本来自 Berlin 情感语音库，采集的情感视频短片也提供了一些测试样本。

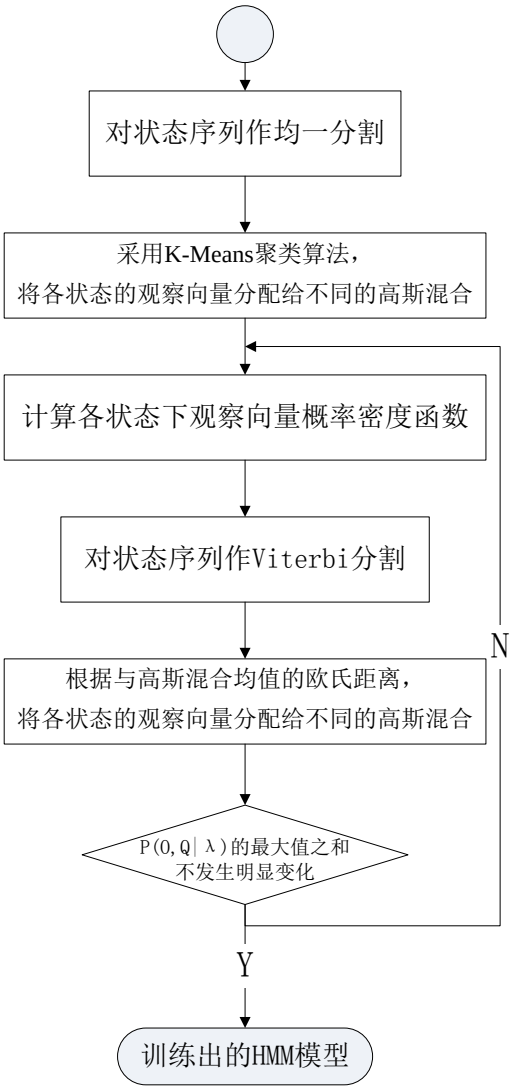

图 5-12 HMM 模型的训练流程

5.3.3 多模态情感识别实验平台

在 VC++.NET 环境下搭建多模态融合情感识别实验平台。实验平台的核心包括表情和情感语音特征提取模块以及融合表情和语音的情感识别模块。

特征提取模块的类按层次结构设计，分为数据模型层、数据访问层、业务逻辑层和控制层。

数据模型层的模板类 Data <Type, dataLength>用于记录面部特征点、表情特征和情感语音特征。Type 包括 double 和 CvPoint 两种，dataLength 包括 aamLength、fapLength、mfccLength、plccLength、pitchLength、formantLength 和 prosodyLength 七种。

数据访问层的抽象基类为模板类 DataAccess <Type, dataLength>, 模板类 FileDataAccess<Type, dataLength>继承自模板类 DataAccess <Type, dataLength>,

用于读写表情特征文件和情感语音特征文件，模板类 DirectoryDataAccess <Type, dataLength>则用于读写情感语音特征文件。

业务逻辑层的类 FAP 用于计算表情特征，类 PROSODY 用于计算情感语音特征。

控制层的类 AAMControl、类 FAPControl 和类 PROSODYControl 调用业务逻辑层的接口，对表情库和语音库的文件作批量处理。

情感识别模块的类也按照与特征提取模块的四层结构设计。

数据模型层的类 HMMParams 提供隐马尔可夫模型的结构参数，类 CvEHMM 用于记录训练的隐马尔可夫模型，类 HMMData 记录计算出的表情特征和情感语音特征关于各种情感状态隐马尔可夫模型的条件概率，类 CvANN_MLP 记录训练的多层感知器。

数据访问层的类 HMMParamsAccess 用于读写记录隐马尔可夫模型结构参数的文件，模板类 HMMAccess <dataLength>用于读写隐马尔可夫模型的 xml 文件，类 HMMDataAccess 用于读写记录条件概率值的文件，类 ANNAccess 用于读写记录多层感知器的 xml 文件。

业务逻辑层的模板类 HMM <dataLength>用于训练隐马尔可夫模型和计算条件概率值，类 ANN 提供函数训练多层感知器。

控制层的类 HMMControl 调用业务逻辑层提供的接口。

5.3.4 隐马尔可夫模型和多层感知器的函数实现

下面是若干个与隐马尔可夫模型的训练和识别算法相关的函数。二维嵌入式隐马尔可夫模型分为两层，第一层的每个状态都代表一个一维隐马尔可夫模型。设置第一层状态数为 1，可以将二维隐马尔可夫模型转化成一维隐马尔可夫模型。

1. 函数 cvUniformImgSegm(CvImgObsInfo* obsInfo, CvEHMM* hmm)对观察向量序列的状态序列作均一分割，无返回值，参数包括隐马尔可夫模型的结构指针和观察向量序列指针。

2. 函数 cvInitMixSegm(CvImgObsInfo** obsInfoArray, int numImg, CvEHMM* hmm)采用 K-Means 聚类算法，将各状态的观察向量分配给不同的高斯混合，无返回值，参数包括观察向量序列指针数组、观察向量序列数和隐马尔可夫模型的结构指针。

3. 函数 cvEstimateHMMStateParams(CvImgObsInfo** obsInfoArray, int numImg, CvEHMM* hmm)根据各状态的高斯混合分割结果，计算状态包含的各高斯混合的权重、均值向量和协方差矩阵，无返回值，参数包括观察向量序列指针数组、观察向量序列数和隐马尔可夫模型的结构指针。

4. 函数 cvEstimateTransProb(CvImgObsInfo** obsInfoArray, int numImg,

CvEHMM* hmm)计算转移概率矩阵，无返回值，参数包括观察向量序列指针数组、观察向量序列数和隐马尔可夫模型的结构指针。

5. 函数 cvEstimateObsProb(CvImgObsInfo* obsInfo, CvEHMM* hmm)根据状态包含的各高斯混合的权重、均值向量和协方差矩阵，计算各状态下观察向量混合高斯分布的概率密度函数，无返回值，参数包括观察向量序列指针数组和隐马尔可夫模型的结构指针。

6. 函数 cvEViterbi(CvImgObsInfo* obsInfo, CvEHMM* hmm)对观察向量序列的状态序列作 Viterbi 分割，求出观察向量序列最可能对应的状态序列，返回类型为浮点型，记录了观察向量序列对应最可能状态序列的概率，参数包括观察向量序列指针和隐马尔可夫模型的结构指针。

7. 函数 cvMixSegmL2(CvImgObsInfo** obsInfoArray, int numImg, CvEHMM* hmm)根据观察向量到各高斯混合均值的欧氏距离，重新将各状态的观察向量分配给不同的高斯混合，无返回值，参数包括观察向量序列指针数组、观察向量序列数和隐马尔可夫模型的结构指针。

多层感知器的训练算法由 CvANN_MLP 类实现，主要包括如下的函数。

1. 函数 create(const CvMat *_layer_sizes, int _activ_func = SIGMOID_SYM, double _f_param1 = 0, double _f_param2 = 0)用于创建多层感知器的拓扑结构，无返回值。

参数_layer_sizes 记录输入层、各隐藏层和输出层的神经元数；_activ_func 表示激活函数，包括 CvANN_MLP::IDENTITY、CvANN_MLP::SIGMOID_SYM 和 CvANN_MLP::GAUSSIAN 三种，其中 CvANN_MLP::IDENTITY 代表单位函数 $f(x)=x$，CvANN_MLP::SIGMOID_SYM 代表 Sigmoid 函数 $f(x)=\frac{\beta(1-e^{-\alpha x})}{1+e^{-\alpha x}}$，CvANN_MLP::GAUSSIAN 代表高斯函数 $f(x)=\beta*e^{-\alpha x*x}$；_f_param1、_f_param2 是激活函数的自由参数。所有神经元有相同的激活函数，且带有相同的自由参数 α 和 β.

2. 函数 CvANN_MLP::train(const CvMat *_inputs, const CvMat *_outputs, const CvMat *_sample_weights, const CvMat *_sample_idx = 0, CvANN_MLP_Train _params = CvANN_MLP_TrainParams(), int flags = 0)实现了多层感知器的训练算法，返回类型为整型。

参数_inputs 表示训练集的输入向量矩阵，每行代表一个输入向量；_outputs 表示训练集的输出向量矩阵，每行代表一个输出向量；_sample_weights 代表权重矩阵和偏置项。

5.4 实验结果及分析

实验测试的情感状态有四种，分别是高兴、悲伤、愤怒和厌恶。实验采用CK库中的图像序列和Berlin情感语音库中的语音信号，通过五折交叉验证，得到的结果如表5-1所示。由表可知，融合表情和语音的多模态算法用于识别情感具有较高的识别精度，达到91.2%。

表5-1 多模态融合的情感识别率

高兴	悲伤	愤怒	厌恶	平均识别率
95.2%	92.9%	93.3%	80.6%	91.2%

表5-2为文献[15]中各种表情的识别率，其平均识别率为84.0%，对比本实验得到的识别结果，融合表情和语音的多模态算法的识别精度高于仅利用表情模态时的精度。

表5-2 文献[15]中的表情识别率

高兴	悲伤	愤怒	厌恶	平均识别率
93.3%	59.6%	72.7%	97.3%	84.0%

表5-3为文献[20]中各种语音情感的识别率，其平均识别率为87.0%，对比本实验得到的识别结果可知，多模态情感识别算法的识别率高于仅利用语音模态的结果。

表5-3 文献[20]中的语音情感识别率

高兴	悲伤	愤怒	惊讶	平均识别率
84.0%	95.0%	85.0%	83.0%	87.0%

5.5 实验结论

实验表明，融合表情和语音的情感识别算法在识别样本中的高兴、悲伤、愤怒、厌恶等情感状态时具有较高的准确率。

从表情图像序列中提取的FAP参数特征对各种不同的表情具有较好的判别性；根据语音信号计算出的短时平均能量、基音频率和共振峰，也能够较好地区分多种情感状态。

多模态情感识别算法的识别结果相比于仅利用语音模态的识别结果有一定的提升，相比于表情模态的识别结果也有一定改进。

第六章 总结与展望

6.1 工作总结

本文提出一种多模态融合的情感识别算法，从面部图像序列和语音信号中提取表情和情感语音特征，基于隐马尔可夫模型和多层感知器设计融合表情和语音模态的情感分类器。

采用面部动画参数作为面部图像序列的表情特征，面部动画参数的计算基于面部特征点在图像序列中的位移，利用主动外观模型可以实现面部特征点的定位和跟踪。对语音信号作时域和频域分析，提取各帧的短时平均能量、基音频率和共振峰作为语音特征。

从训练样本中提取的表情和情感语音特征作为观察向量序列，采用 Viterbi 算法训练各种表情和语音情感的隐马尔可夫模型，计算观察向量序列关于各 HMM 模型的条件概率，作为多层感知器的输入；采用反向传播学习算法训练多层感知器的权重矩阵。从测试样本中提取表情和情感语音特征构成观察向量序列，计算观察向量序列关于各 HMM 模型的条件概率，作为多层感知器的输入；计算多层感知器的输出，识别出测试样本的情感状态。

在 VC++.NET 环境下搭建实验平台；采用 Cohn-Kanade2010 人脸表情库和 Berlin 情感语音库作为情感识别数据库；实验测试的情感状态有四种，分别是高兴、悲伤、愤怒和厌恶。实验表明，融合表情和语音的情感识别算法在识别样本中的高兴、悲伤、愤怒、厌恶等情感状态时具有较高的准确率；多模态情感识别算法的识别结果相比于仅利用语音模态的识别结果有一定的提升，相比于表情模态的识别结果也有一定改进，多模态情感识别算法的准确率高于单模态识别的结果。

6.2 未来展望

情感特征提取是多模态情感识别的关键环节。情感特征的区分性决定了情感状态分类的准确性；特征提取算法的时间复杂性则对情感识别的实时性有重要影响。未来工作中可以考虑改进语音特征的提取，寻找能够更有效地区分各种情感的语音特征；考虑实时性更好的表情特征提取算法，如光流法等。

可以尝试其它的融合表情和情感语音特征的方法，如 D-S 证据理论等，通过比较各种方法的优劣，找到更合适的模态融合算法，设计出识别精度高、实时性好的情感分类器。

情感表达的模态包括面部表情、语音、姿势、生理信号、文字等。可以考虑从面部图像和语音以外的模态中提取情感特征，以提高情感识别的精度。

参考文献

[1] 崔景霞. 小波变换在人脸表情识别研究中的应用[J]. 长春理工大学学报，2011，34(3)：142-145.

[2] 丁志起,赵晖. 结合差图像和 Gabor 小波的人脸表情识别[J]. 计算机应用与软件，2011，28(4)：47-49.

[3] 祝长生，王志良，宋倩霞等. 表情识别的图像预处理和特征提取方法研究[J]. 小型微型计算机系统，2009(6)：1155-1159.

[4] 张杰. 基于弹性模板和 K 近邻结合的表情识别算法[J]. 太原师范学院学报：自然科学版，2011，10(4)：87-90.

[5] 朱明早，罗大庸，王一军. 基于图像重建的表情识别算法[J]. 中国图象图形学报，2010(1)：98-102.

[6] 郑秋梅，吕兴会，时公喜. 基于双向二维直接线性判别分析的人脸表情识别[J]. 中国石油大学学报：自然科学版，2010，34(5)：179-182.

[7] 刘松，应自炉. 基于局部特征和整体特征融合的面部表情识别[J]. 电子技术应用，2005，31(3)：4-6.

[8] 张建明，张晓翠. 一种处理部分遮挡表情图像的方法[J]. 计算机工程与应用，2011，47(3)：170-173.

[9] 周川，林学闾. 基于核函数因素分解模型的表情合成与识别[J]. 清华大学学报：自然科学版，2006，46(10)：1751-1754.

[10] Bassili J. Emotion recognition: the role of facial movement and the relative importance of upper and lower areas of the face[J]. Journal of Personality and Social Psychology，1979，37(11)：2049-2058.

[11] Ekman P，Friesen W. Facial Action Coding System Consulting[M]. Psychologists Press Inc.，California，1978.

[12] Tian Y，Kanade T，Cohn J. Recognizing action units for facial expression analysis[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence，2001，23(2)：97-115.

[13] Pardas M，Bonafonte A，Landabaso J L. Emotion recognition based on MPEG4 facial animation parameters[C]. Proceedings of IEEE Acoustics, Speech and Signal Processing，2002.

[14] Landabaso J，Pardas M，Bonafonte A. HMM recognition of expressions in unrestrained video intervals[C]. Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing，2003(3)：197-200

[15] Pardas M，Bonafonte A. Facial animation parameters extraction and expression recognition using Hidden Markov Models[J]. Signal Processing: Image Communication，2002，17(9)：675-688.

[16] 余棉水，黎绍发. 基于光流的动态人脸表情识别[J]. 微电子学与计算机，2005(7)：113-115.

[17] 杨国亮，于仲安. 基于改进光流算法和 HMM 的面部表情识别[J]. 微计算机信息，2008(1)：284-286.

[18] 丁志起,赵晖. 结合差图像和 Gabor 小波的人脸表情识别[J]. 计算机应用与软件,2011，28(4)：47-49.

[19] 赵力，钱向民. 语音信号中的情感识别研究[J]. 软件学报，2001，12(7)：1050-1055.

[20] 赵力，将春辉，邹采荣等. 语音信号中的情感特征分析和识别的研究[J]. 电子学报，2004，32(4)：606-609.

[21] Lin Y，Wei G. Speech emotion recognition based on HMM and SVM[C]. Proceedings of the Fourth International Conference on Machine Learning and Cybernetics，2005：4898-4901.

[22] 康燕,张雪英. 基于 ZCPA 参数的语音情感识别研究[J]. 山西电子技术,2011(3):80-82.

[23] 屠彬彬，于凤芹. 基于 EMD 的改进 MFCC 的语音情感识别[J]. 计算机工程与应用，2012，48(18)：119-122.

[24] 叶吉祥,张密霞,龚希龄. 基于 MF-DFA 的语音情感识别[J]. 计算机工程与应用,2012，48(18)：119-122.

[25] 余华，黄程韦，张潇丹，金赟，赵力. 混合蛙跳算法神经网络及其在语音情感识别中的应用[J]. 南京理工大学学报：自然科学版，2011，35(5)：659-663.

[26] 张石清，李乐民，赵知劲. 基于一种改进的监督流形学习算法的语音情感识别[J]. 电子与信息学报，2010(11)：2724-2729.

[27] 黄程韦，金赟，王青云等. 基于语音信号与心电信号的多模态情感识别[J]. 东南大学学报：自然科学版，2010，40(5)：895-900.

[28] Zeng Z，Tu J，Brian M，Huang T S et al. Audio-visual affective expression recognition through multistream fused HMM[J]. IEEE Transactions on Multimedia，2008，10(4)：570-577.

[29] Zeng Z，Tu J，Liu M et al. Audio-Visual Affect Recognition[J]. IEEE Transaction on Multimedia，2007，9(2)：424-428.

[30] Zeng Z，Tu J，Pianfetti B. Audio-visual affect recognition through multi-stream fused HMM for HCI[C]. Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition，2005(2)：967-972.

[31] Hoch S，Althoff F，McGlaun G et al. Bimodal fusion of emotional data in an automotive environment[C]. Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing，2005(2)：1085-1088.

[32] Go H J，Kwak K C，Lee D J et al. Emotion recognition from facial image and speech signal[C]. Proceedings of International Conference on Instrument and Control Engineers，2003: 2890-2895.

[33] Cootes T F，Edwards G J，Taylor C J. Active appearance models[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence，2001，23(6)：681-685.

[34] Cootes T F，Edwards G J，Taylor C J. Active appearance models[C]. Proceedings of European Conference on Computer Vision，1998：484-498.

[35] 廉文娟，山世光. 使用 ASM 与 AAM 方法进行器官定位[J]. 山东科技大学学报（自然科学版），2002，21(3)：40-43.

[36] Cootes T F，Taylor C J，Cooper D H et al. Active shape models—their training and application[J]. Computer vision and image understanding，1995，61(1)：38-59.

[37] Cootes T F，Taylor C J，Lanitis A et al. Building and using flexible models incorporating grey-level information[C]. Proceedings of Fourth International Conference on Computer Vision，1993：242-246.

[38] Cootes T F，Taylor C J. Active Shape Model Search Using Local Grey-Level Models: A Quantitative Evaluation[C]. Proceedings of British Machine Vision Conference，1993：639-648.

[39] Lucey P，Cohn J F，Kanade T et al. The Extended Cohn-Kanade Dataset (CK+): A complete dataset for action unit and emotion-specified expression[C]. Proceedings of IEEE workshop on CVPR for Human Communicative Behavior Analysis，2010：94-101.

[40] Nordstrom M M，Larsen M，Sierakowski J el al. The IMM Face Database—An Annotated Dataset of 240 Face Images. Technical report, Informatics and Mathematical Modeling，Technical University of Denmark，2004.

[41] Burkhardt F，Paeschke A，Rolfes M el al. A Database of German Emotional Speech[C]. Proceedings of 9th European Conference on Speech, Communication and Technology，2005：1-4.

发表论文和参加科研情况说明

发表论文：

Chao Xu, Tianyi Cao, Zhiyong Feng, Caichao Dong. Multi-Modal Fusion Emotion Recognition Based on HMM and ANN. International Conference on E-business Technology and Strategy, 2012, pp 541-550 (Corresponding author: Tianyi Cao)

致　谢

首先要感谢王建荣老师，王老师严谨的治学态度和科学的工作方法给了我很大的帮助和影响。

本论文的完成离不开冯志勇教授的悉心指导，冯老师不但让我在技术水平上得到了很大的提高，更教会了我该如何独立解决问题、如何在团队中寻求共同进步，这些将使我受益终生。在此衷心感谢冯老师对我的关心和指导。

在实验室工作及撰写论文期间，还得到了许多老师和同学的帮助。特别要感谢的是徐超老师，在研究工作和论文写作期间给我了许多具体的指导，在此表示由衷的感谢。

我感谢实验室的师兄师姐师弟师妹对我的关心和帮助，感谢梁景莲、李超、董彩超、王丹丹、薛万利、张东萍、陆泽萍、彭伟龙、陈鑫、李超、尹晓燕、常方媛、吕亚丹。

我感谢室友对我的关心和帮助，感谢黄兆桐、沈敏、余小飞。

我感谢我的父母。

师长与同学在工程实现与论文编写过程中对我提供的帮助不仅于此，但本文篇幅有限，未能一一列出，在此一并感谢。